**Trust Management in Social Networks:**

**Fake Account Detection**

**and**

**Analysis**

**Group 3**

**GROUP MEMBERS**

|  |  |  |
| --- | --- | --- |
| **Name** | **ASU ID** | **Email ID** |
| Matthew Dexheimer | 1203066132 | msdexhei@asu.edu |
| Shruti Sonawane | 1209340908 | ssonawan@asu.edu |
| Sruthi Maddineni | 1208647085 | smaddin2@asu.edu |
| Lydia Manikonda | 1203979369 | lmanikon@asu.edu |
| Keerthi Sree Koduri | 1206341834 | kkoduri1@asu.edu |
| Runyu Jin | 1209995303 | rjin9@asu.edu |
| Sitanshu Mishra | 1206136577 | smishr16@asu.edu |
| Sai Indraneel Patcha | 1207563925 | spatcha@asu.edu |
| Kandhan Sekar | 1209331327 | ksekar1@asu.edu |
| Sooraj Raja Gopal | 1210315246 | srgopal@asu.edu |
| Prashanth Artal | 1207627157 | partal@asu.edu |

**TABLE OF CONTENTS**

**1. Summary**

**2. Introduction**

**Motivation and background**

**Goals and scope of study of the project**

**3. Overview**

**4. Detailed Results**

**Approach**

**Coverage**

**Data & Analysis**

**Strengths and Weaknesses**

**5. Conclusions & Recommendations**

**Comparison of Strengths and Weaknesses**

**Future Work**

**6. Evaluation Questions**

**7. References**

**SUMMARY:**

Characterize and understand trust management in social networks by researching specific topics related to fake accounts and networks, factors leading to their success, and exploring different ways in which their deception works.

Social networks are growing at a fast pace that is leading to new challenges in terms of preserving and protecting the end users’ privacy and the credibility of active user accounts. These challenges subsume different questions such as how much credence we give to each source or each user on the network, how different social interactions indicate friendship, support or approval and at the same time how they signify the disapproval or distrust of the opinions of others.

Users increasingly rely on the trustworthiness of the information exposed on Online Social Networks (OSNs). In addition, OSN providers base their business models on the marketability of this information. However, OSNs suffer from abuse in the form of the creation of fake accounts, which do not correspond to real humans.

**INTRODUCTION:**

**Motivation and background**

Social network is a map of relevant connections which can be similar interests, hobbies, activities, relationships, etc., between individuals, organizations, countries, etc. This term “social network” was first coined by Professor J. A. Barnes in the 1960s [Barnes 1967] describing the definition of social network as the network of associations formed for support such as emotional, instrumental, appraisal and information. With the rampant advancements in technology, these social networks are formed online and play an important role in establishing connections between individuals and organizations and bringing them more closer. In the recent years, online social networks like Facebook, Twitter, Instagram have grown exponentially where users or individuals or organizations join these platforms and establish connections. There are different motivations of why users join these online platforms and these range from sharing their updates, connecting with friends and other users who share similar motivations. In this scenario, trust is an important component that decides whether to a connection will be made between two users or organizations on these online social networks.

Trust has been well studied and defined differently in multiple fields of science for example, Psychology, Philosophy, Sociology, as well as Computer Science. Trust is defined based on the context and stands differently based on the user or individual you consider. In one field trust is the “subjective expectation an agent has about another’s future behavior based on their history of encounters” [Mui et al., 2002]. Trust also has multiple properties such as :

1. trust is asymmetric which means in a scenario of two users U1 and U2, U1 trusts U2 100% but U2 trusts U1 only 20%;
2. trust can be transitive which is a very important property on online social networks where in a scenario of three users U1, U2 and U3, U1 trusts U2, U2 trusts U3 and since U1 and U3 has one mutual friend or connection, U1 trusts U3.

There are multiple ways that the trust is maintained in online social networks however, trust is misused when fake accounts are created for malicious purposes. Fake accounts are accounts that mislead users online and they are created online to serve fraudulent purposes. Literature [6,7,12] revealed that fake accounts are usually created for beneficial malicious activities like spamming, click-fraud, identity theft, etc. In contrast, fake accounts are also created for social purposes [9] like friendly pranks, concealing their real identity to maintain privacy, online social games, etc. For these fake accounts to be thrive online, these accounts have to gain trust from other users with whom they want to establish connection.

In this paper, we focus on understanding how these fake accounts are earning trust from genuine users and prospering online. Studying fake accounts is very important to understand multiple ways how security is breached online and proposing early interventions to protect genuine users online. Past literature [10] also suggests that if information flow from or through fake accounts is hindered on time, it can help reduce the damage it is aiming for by almost 75%. In this paper we focus on these multiple aspects of detecting and understanding different factors that play a key role in characterizing the activities of fake accounts and we also focus on how the information flow is different on these networks compared to the real social networks.

**GOALS AND SCOPE OF STUDY OF THE PROJECT**

**Detecting or identifying fake accounts & predicting friends vs foes**

Our goal is to determine how fake accounts can be detected. Attackers may create a fake account. The attacker may use this fake account to gain users personal information. This potential attack increases legitimate users’ risk. Therefore, the detection of fake accounts will be important. First, we must be able to identify fake accounts. We must determine what systems are currently in use. We must determine if those systems are sufficient. Also, we must determine if there are better approaches.

**Factors leading to the success of fake accounts**

Our goal is to study the attributes leading to the success of fake accounts in Online Social Networks. Learning more about the attributes that aid fake account holders help OSNs to design techniques to detect fake accounts or spammers and making the creation of fake accounts difficult for attackers. Studying these attributes help OSNs to protect the privacy of users, integrity of the data available on OSNs and inhibit fraudulent activities. The scope of the discussion is limited to general methods used by popular websites, their effectiveness in managing fake users and the loopholes in these measures that are exploited by fake accounts. The future scope lists some effective proposals to be implemented.

**Studying the network and behavior of fake accounts**

Any large-scale online social network system faces constant security threats through the presence of malicious entities. When such entities are each capable of presenting a large network of identities in a social network, then a substantial control of the system gets compromised, thereby making any redundancy techniques useless. Similarly these fake accounts may be spread across multiple social networks as well. There are many detection methods employed to handle such fake identities, but the scale and timeliness of these methods is still a concern.

Also Studying the behavior of the fake account users is really important in distinguishing them from the legit users. We represent different case studies and methods to understand the behavior of fake accounts.

**How do people fall for fake accounts**

The proliferation of fake accounts is costly to both legitimate users and the OSNs. By becoming aware of the dangers on OSNs, and employing smarter strategies both users and networks can protect themselves from potential threats. This project primarily serves as an in-depth study into the nature of OSNs and what causes legitimate users to fall prey to fake accounts. Our goal is to identify approaches researchers have taken to discovering whether an account is real or fake, and to utilize that information in providing some additional strategies to users for protecting themselves online. Our scope of the project include the following research topics.

**Information diffusion in fake accounts compared to legitimate accounts**

Social media plays a key role in the formation and dissemination of information in real-time. There are multiple ways that an information reaches users and two main ways are through connections on these networks and through external or offline news sources which will later be disseminated on the network. The type of information getting diffused should be trustworthy as diffusion occurs rampantly and affects significant number of users or nodes in the networks which may lead to disastrous scenarios if the news is fake or not trustworthy. In this regard, we study and investigate how the diffusion of fake news and diffusion on untrustworthy users can be modeled is very important. In this paper, we explore multiple diffusion models used in different real-world scenarios on multiple popular online social platforms like Twitter and the lessons we learned that should be considered when developing new diffusion models.

**OVERVIEW:**

**Matthew Dexheimer:**

I am in the subgroup in charge of researching how fake accounts are detected. I began by researching what methods of fake account detection are currently implemented. I have also researched the effectiveness of currently implemented fake account detection system.

While researching the papers assigned to me, I helped organize my subgroup and facilitated collaboration between team members. My subgroup discussed the papers we had read and the implications of them. Through these discussions, I learned about the current methods of fake account detection and other, potentially more effective methods of fake account detection.

With regards to the writing the current paper, I helped organize meetings, make decisions regarding strategy, and write passages of the paper. This role has allowed me to gain experience working with large groups.

This assignment has allowed me to gain a better understanding of how trust management for OSNs work, especially with regards to fake accounts.

**Shruti Sonawane:**

The topic of my research was detecting fake as well as compromised accounts in Online Social Networks (OSNs). For the detection of fake accounts, I studied a paper based on a system namely SybilRank that does not follow the traditional machine learning based approach to detect fake accounts in an OSN. SybilRank system produces a ranked list of accounts that are highly likely to be fake and the ones that are least likely to be fake. The approach describes its statistics based on an experimental implementation in Tuenti, which is Spain’s largest OSN and the experimental results prove more efficient and promising as compared to earlier methods, as the SybilRank system is cost effective even for large OSNs like Facebook or Twitter. Studying this paper gave me an insight into how complex social graphs can be and the task of detecting fake accounts is truly beyond the scope of manual effort.

I also studied a paper that concentrated on detection of compromised accounts. I understood the clear difference between a fake account and a compromised account. A compromised account is basically a legitimate account that was taken over by a malicious attacker. The paper also helped me understand that such compromised accounts will pass undetected through fake account detection softwares and what needs to be done to specifically deal with these accounts.

Overall, studying these papers were a great learning experience for me in the field of social network security and also from the point of view of collaborating with a team.

**Prashanth Artal:**

My research basically focused on the detection of fake accounts in Online social network. During research I came across different approaches that the authors have followed for detecting the fake accounts.Then, I compared the approaches and among them I found an approach which is best compared to that of any other approach. As this approach mainly deals with the detection of fake accounts at cluster level so the time taken for the detection is totally reduced. Even though they have used different classification algorithms but they were able to explain the best classification algorithm for the experiment. And in another approach, the authors were able to come up with minimum number of attributes and able to get maximum accuracy of spammers. But this approach is applicable with respect to only twitter and the detection happens at account level so this approach takes more time compared to cluster level detection.

**Sruthi Maddineni:**

I studied about different attributes that lead to the success of fake accounts in Online Social Networks. Fake accounts mislead and manipulate the social media by misreporting, spreading rumors and create chaos or just spam the social media by creating noise. Anyone can create an OSN account by providing a valid email address. Attackers create profiles to attract as many users as possible and build their network. The attackers exploit the trust established with other users in their network and steal private information of victims and use it for their benefit. There are several resources available online which are being exploited by attackers to carry their attacks on victims in OSNs, most of which are inexpensive. We need to study about different attributes to develop defences or countermeasures against these attacks and help protecting the privacy of OSN users and the integrity of the data available on OSNs.

**Lydia Manikonda:**

I investigated the subtopic focusing about how information diffuses through fake accounts and also how fake or non-trustworthy news in general diffuse faster. I initially started with a seed set of papers that are very well cited on this particular topic on Google Scholar and using the references mentioned in these initial set of papers, I conducted an in-depth analysis starting from the general overview of how diffusion happens on online social networks and there by focused on how diffusion happens through fake users online. Through my analysis, I have learned that the current available literature heavily focuses on investigating how fake news spreads and the reasons why it spreads so fast. Also, in my in-depth research I have read multiple case studies with Twitter as the main social media platform which analyzed the various parameters and constraints that lead to the formation of fake news and how this news disseminates online. The lessons learned from conducting the research paper survey highlighted the importance of protecting and defending the information and information systems which in our case happen to be user information and the social networks to ensure confidentiality, integrity, availability and non-repudiation which are the main fundamentals of information assurance. We as subgroup are focusing on submitting this work as a survey paper in a journal on how the current state-of-the-art literature focuses on how information diffusion happens when the information is fake and how fake users spread information which is not trustworthy.

**Keerthi Sree Koduri:**

My subtopic is the factors leading to the success of fake accounts. I have divided it into two other topics in the context of fake account generation and fake account usage. I studied about the mechanisms using which fake accounts can be successfully created. I have researched about the various tools that are used to automate the process of fake account generation and their capabilities to handle challenges on the popular websites.

I have also gone through the actual study performed on the most popular websites and the measures they have taken to counteract fake accounts. Next, I have compared various measures and check which website is more prone to attacks. Then I have read about the social engineering techniques using which a fake user can steal information.

In the end, I have discussed about the future scope of the work that must be done to stop fake accounts.

**Runyu Jin:**

I am in the subgroup of ‘How do people fall for fake accounts’. I read papers on this part and analysed different reasons on people fall for fake accounts.We compared different approaches on this used in different papers and made some analysis on the results. During the process, I acted actively to discuss papers with my teammate and exchange our ideas to have more inspirations. Sometimes the papers may be difficult to understand, I will go to my teammate for help.

I am also the team leader of the whole team. As a team leader, I organized some meetings and during the meeting, I encouraged team members to give out ideas actively and wrote some meeting notes. I also divided workload among group members and make all the things organized and in good order. I really learned a lot from this class, such as how to get useful information from papers and how to deal with team members for better cooperation and most importantly how to be a team leader.

**Sitanshu Mishra:**

As part of the sub-group for How do People Fall for Fake Accounts, I researched the topic by reading academic research papers from reputable institutions such Association of Computer Machinery and Institute of Electrical and Electronics Engineers. Through active reading, researching, and discussions both within the sub-group and the group as a whole, I was able to learn about an extensive topic and formulate my own thoughts and opinions on the different approaches and techniques studied.

In terms of the final deliverable, I worked actively within my subgroup and group to remain focused on my specific part without overlapping. Worked towards facilitating and incorporating the research of various different subgroups into one cohesive unit, through proofreading and maintaining the same tone throughout the paper.

**Sai Indraneel Patcha:**

As we mentioned in our scope of study, that we go along with the research topics. My area of concentration is on the Research Topic – 2(RT-2) i.e. Studying the network/ behavior of the fake account users in online social networks.

I started reading several papers and picked some of the best that are relevant to my research topic. I started with understanding the behavior of fake account users and how they get along in online social networks, to get a thorough knowledge of how the fake account users can be distinguished from the legit users. In this paper, I started with a case study on Pengyou(One of the popular social networking site) that shows how fake account users got the important information from legit users , to understand the network of fake account users. Later on I moved to understanding the different methods, which shows the behavior of fake account users. One of the methods is to detect identity deception using non verbal behavior in Wikipedia, that can be applied to any online social network. Then I moved to other method which shows the anomalous behavior in online social networks using PCA (Principal Component Analysis).

**Kandhan Sekar:**

My research was primarily focused on Information diffusion of untrustworthy news or fake news, wherein the papers i had studied were based on how information diffuses in social network, taking Twitter as a case study. In the twitter example, the research revealed some of the threats and negative impacts of information diffusion such as accounts hijacking, rumor and non-rumor tweets, spam URL and botnet detection. It paved the way to understand the strategies used by spammers, hackers, botnets to misuse, hijack accounts or tweets to acquire information or spread unwanted information. How this loss of information affects the user, the social network and the efficiency of strategies currently used to detect and prevent were also investigated. I realized the need for information assurance in social networks and how its very implementation is impacting social networks at large and why it needs to be improved and how can we achieve that.

**Sooraj Raja Gopal:**

Within our group project, my responsibility was to go in-depth on the subtopic “Studying the network of fake accounts”, for which I had to consider material on how fake account groups are used by an adversary to launch attacks on Online Social Networks. I have read through and analysed relevant research papers and have used them to summarize and detail out the various methods and approaches whereby such attacks come about, and how to actively detect them. The different approaches have been compared and strengths and weaknesses detailed out as well. Overall it has been very insightful in understanding how to better manage one self’s information on online networks and how to be better equipped to protect our accounts against adversaries.

**DETAILED RESULTS:**

**How to detect or identify fake accounts and predict friends vs. foes.**

**Approach:**

The Facebook Immune System is Facebook’s defense against malicious users, including fake accounts. A paper published in *Proceedings of the 4th Workshop on Social Network Systems* in 2011 by Tao Stein, Erdong Chen, and Karan Mangla outline how the Facebook Immune System works.[48] They outline the adversarial cycle and how their system decreases the time that the attacker is in control while prolonging the amount of time that the defender is in control.

The adversarial cycle indicates the states that the OSN can be in.[48] It is divided into two major parts: adversary control and defender control. The adversary control begins with an attack. The attack causes the attacker to gain control. The attack continues until the defender detects the attack. At this point the defender regains control by activating their defenses. The defender retains control until the attacker mutates their strategy and launches a new attack.

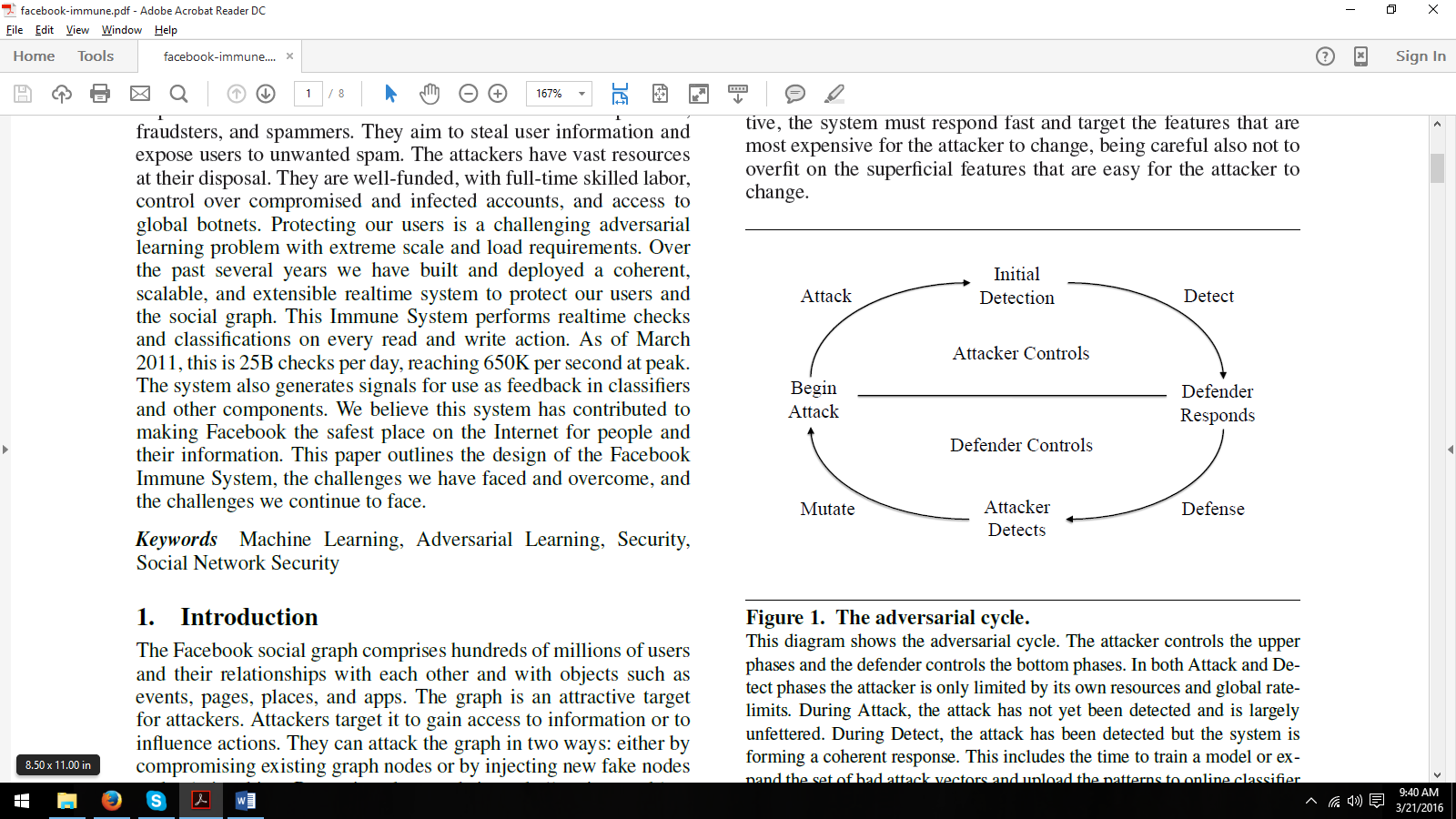


Figure 1: This figure shows the adversarial cycle laid out in the description of the Facebook Immune System.[48]

**Coverage:**

The Facebook Immune System consists of five different parts.[48] First, the classifier service is designed to classify user actions using different machine learning algorithms. Second, the feature extraction language (FXL) is a dynamically executing functional language designed to facilitate the classifier and policy engines. It is Turing-complete and statically typed. Third, dynamic model loading allows models build by the feature extraction to be into the classifier service without the system restarting. Fourth, the policy engine allows Facebook to control the policies being implemented. Fifth, feature loops aggregate data from user feedback, data crawlers, and query data.

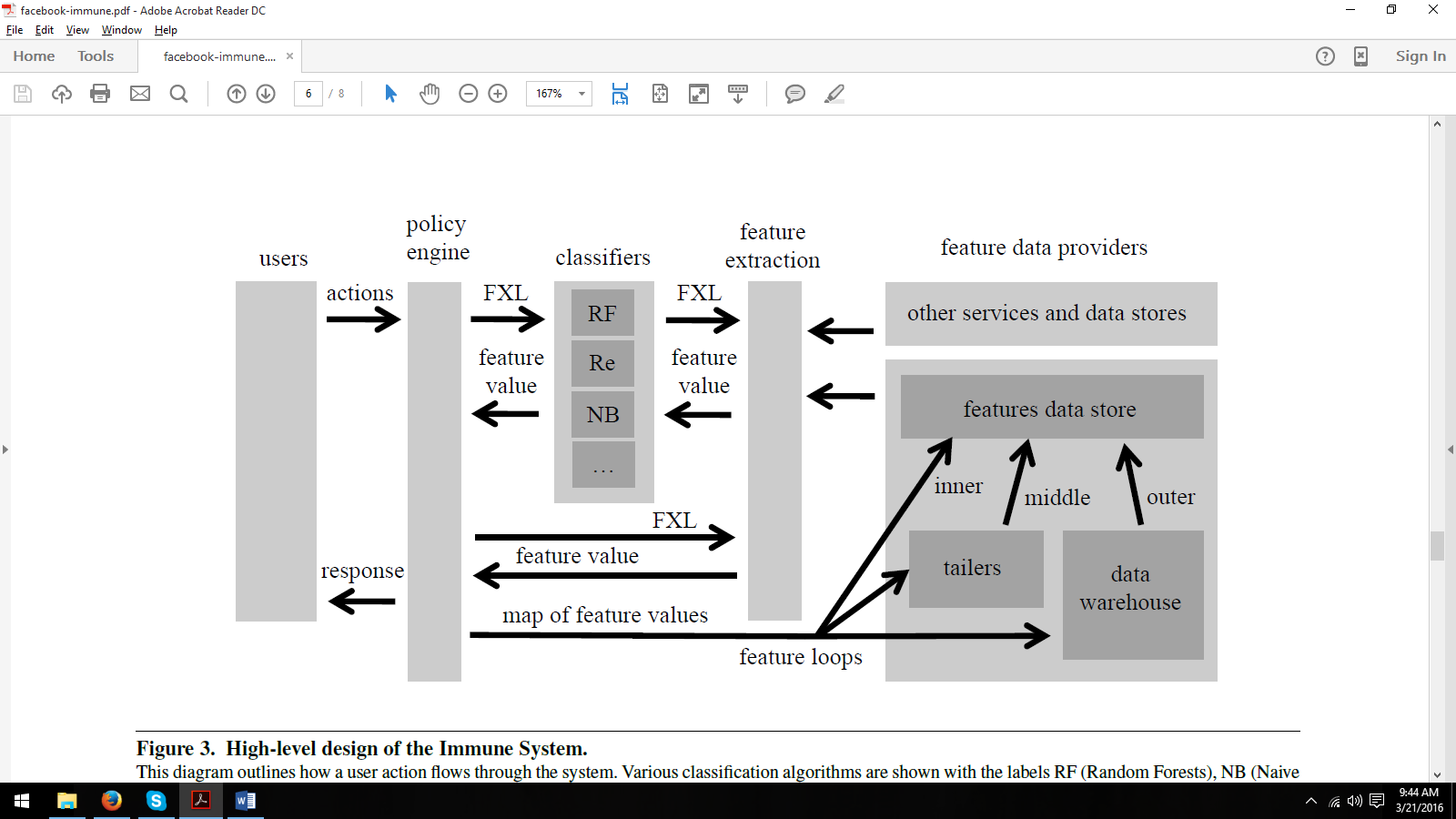


Figure 2: This shows the conceptual layout of the Facebook Immune System.[48]

FXL is a prominent part of the Facebook Immune System. Its creation was motivated from a need to constantly update due to the dynamic nature of attacks.[48] FXL is designed to make it easy to test new ideas quickly. It allows for features to be formed by combining other features. FXL was chosen to be a functional language because it allows for expressions to easily be represented as trees, allowing for ease of parallelization.

The feature loops are another major part of the Facebook Immune System. They consist of three different types: inner, middle, and outer.[48] The inner loop is in charge of incrementing and decrementing counters tracking features such as the number of times that a link has been posted in the past hour. The middle loop applies feature operations on a larger data set than only counters and has access to a larger database than the inner loops. The outer loop is used to examine potentially harmful features such as a link being posted a large number of times by a small number of users.

**Experimental Results and Conclusion:**

This paper did not report any tests or experiments conducted on the Facebook Immune System.

**Strengths and Weaknesses:**

The authors posit that the Facebook Immune System will be sufficient to meet the challenges of detecting fake accounts on Facebook. This claim is not supported by experimental data.

**Approach:**

One way to help identify fake accounts is to create your own fake account. A paper published in *Proceedings of the 27th Annual Computer Security Applications Conference* in 2011 by Boshmaf et al. reports on an attempt to create a network of fake accounts.[3] In this paper, they create their own network of fake accounts on Facebook controlled by an automated system. The paper lays out the ways in which certain obstacles to creating a fake account can be overcome. Knowing these techniques can be useful for identifying fake accounts.

**Coverage:**

The first obstacle discussed was the CAPTCHA system. The solution used by the authors is a case where it is easier to circumvent a problem, than it is to solve it. In order to solve the CAPTCHA, the authors would have had to create an algorithm to read the CAPTCHA image and produce the corresponding text. However, what the authors did was circumvent creating an algorithm by outsourcing the problem to humans.[3] The fact that the CAPTCHA system can be circumvented indicates that it is insufficient to guard against fake OSN accounts.

The next obstacle discussed was the account creation system. While trivial to a human, it poses an obstacle to an automated system.[3] However, the authors again find a solution. For an e-mail address, the authors use mail.ru, a free e-mail provider to create as many e-mail accounts as needed. The authors used the website hotornot.com to gather all profile pictures.

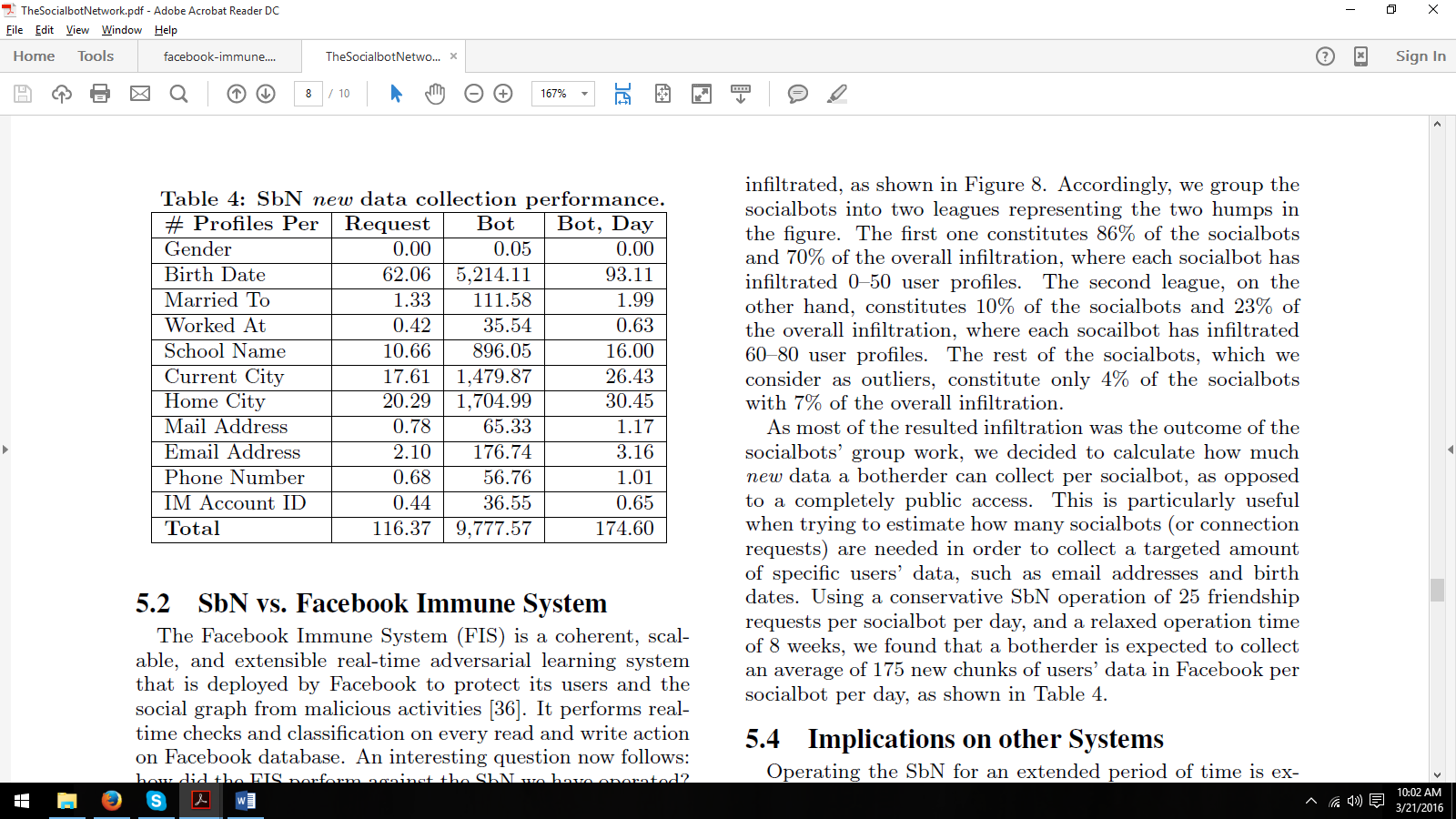


Table 1: This table shows the rate of personal informational retrieval by the socialbot system.[3]

**Experimental Results and Conclusion:**

Out of the 102 fake accounts operating for 8 weeks, only 20 were blocked by the Facebook Immune System.[3] These accounts were flagged as spam, rather than fake accounts. Furthermore, these accounts were flagged by users, rather than the Facebook Immune System. Of the overall attack 70% of fake accounts succeeded in connecting to 0-50 real accounts. In addition 23% of fake accounts succeeded in connecting to 60-80 real accounts. The system was able to collect an average of 175 chunks of new data every day. These chunks were personal identification information such as home city and birth date.

**Strengths and Weaknesses:**

Results of the paper use statistical analysis of experimental data to show that the current system, Facebook Immune System, is insufficient. The method used is successful at circumventing the Facebook Immune System.

**Approach:**

Another approach to detect fake accounts is published in paper “Fake Account Detection in Twitter Based on Minimum Weighted Feature set” by Hashmen Hefny, Ahmed El Azab, Amira M Idrees and Mahmoud A Mahmoud.They followed an approach by checking the behavior of person such as friends in his profiles, number of tweets per day, retweets etc. by making an assumption that the actual legitimate account holders behavior is different from the fake accounts, so exploiting this behavior might lead to detection of fake accounts [49]. For detection, the attributes needs to be considered, in the paper[50] they have considered an attributes of 23 and have reached an accuracy of 84.5% as spammers, but in the present paper, by considering small set of attributes they were able to achieve more accuracy in detection of fake accounts[1].

**Coverage:**

The approach which is mentioned in the paper consists of two main steps, the first step is finding the factors which influences the detection fake accounts and the second step is applying the classification algorithm which uses the factors which are generated in the first step over the twitter profile accounts for detecting the fake accounts[1]. As mentioned earlier in the approach section, the proposed method uses minimum number of attributes which helps in detecting the fake accounts with the great accuracy. Moreover, finding the minimum set of features for extracting, preparing and analyzing is considered to be most effective directions for the detection of fake accounts with maximum accuracy[1]. The measurement of the accuracy is mentioned below in this analysis of experimental data section. The number of features considered in this experiment is 19.

**Dataset:**

The dataset used for this experiment is collected from “the Fake Project”[51]. The researchers who carried out this experiment contacted the authors of “the fake project”, the authors mentioned that they have collected the dataset from different sources such as #elezioni2013 which is verified that it has 1481 accounts which belongs humans. So for fake accounts, in [52] has mentioned about 1000 fake accounts available in http://fastfollowers.com,http://intertwitter.com and http://twittertechnology.com.

**Weighted Features Selection Step:** The researchers have collected all the features from the earlier papers and applied the GAIN Measure on the training dataset for producing weight for all the features.

**Classification Algorithm Selection Step:** For the experiment they have five classification algorithms they are Random Forest[53][56],Naïve Bayes[55], SVM[54], Decision Tree[57] and Neural Network[58]. The results generated by each algorithm is summarized by standard indicators: True Negative(TN), True Positive(TP),False Negative(FN) and False Positive(FP)

**Summary of Experimental Data:**

The experiments are conducted in three steps, the first step is the discussion of five algorithms which applied on the attributes

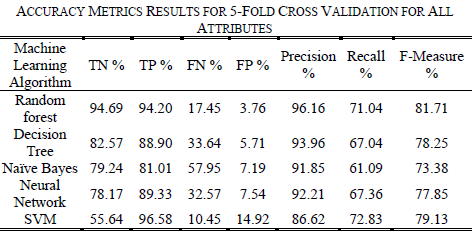


Table 2: Accuracy Metrics results for 5-fold cross validation for all attributes[1]

and second step is the discussion of five algorithms applied on weighted attributes using GAIN Measures and in

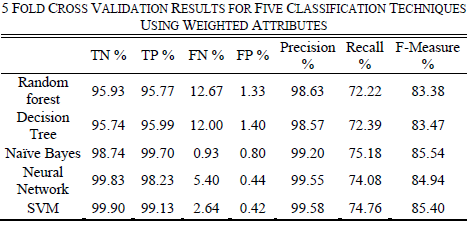


Table 3: 5 Fold cross validation results for five classification techniques using weighted Attributes[1]

the last step is the discussion of five algorithms on the attributes which is above 50%.

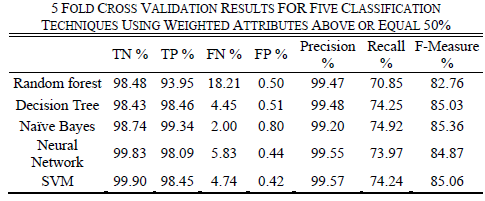


Table 4: 5 FOld cross validation results for five classification using weighted attributes above 50%[1]

**Strength and weakness:**

1. The proposed method uses less number of attributes for the detection of fake accounts with more accuracy compared any other earlier methods.
2. But the weakness of the proposed method is they are using five classification algorithms for the experiment and they have not pointed out one specific algorithm.

**Approach:**

Yet another approach which is efficient than earlier approach is presented in a paper ‘*Detecting clusters of fake accounts in Online Social Networks’* by Cao Xiao, David Mandell Freeman and Theodore Hwa, they have come up with **time sensitive machine learning algorithm** to detect group of fake accounts registered by same person. And this approach solves the problems involved in the earlier approach i.e., instead of detecting each account is fake or not, they have used the method of creating clusters of account and detecting the fake accounts in clusters based on the result weight given by the machine learning algorithm[2]. They have divided the approach into two steps, in the first step they group accounts the accounts into clusters based on the criteria and in the second step the machine level algorithm takes clusters as input along with features. These features are described for whole cluster rather than individual accounts and this approach is scalable to online social networks as numerous profiles will be created in a day so detection of clusters of fake accounts reduces the time taken when compared with the detecting account by account in an Online social network[2]. The machine learning algorithm use the features which are available at the time of registration of profile and shortly thereafter this means the approach detects the profile during the registration itself it means this method is not allowing any attacker to enter the OSN and he/she is being removed during the time of registration.

**Coverage :**

The designed and implemented machine learning pipeline which involves data pre-processing, feature extraction, prediction of clusters of fake accounts and validation. Pipeline consists of major components such as cluster builder, profile featurizer and account scorer[2].

**1] Cluster Builder:** This is the first step of pipeline, where it takes raw list of accounts and builds the clusters of accounts along with features. In order to create clusters of accounts then it should be created based on a criteria so this module takes user defined input such as Max and Min of size of clusters or time duration the accounts are registered and clustering criteria[2]. Once the clusters are built the using user defined criteria can be used to filter some clusters which are not likely to be suspicious. The input for the cluster builder is the raw tables of profiles and it outputs the table of profiles along with the features which can be used for feature engineering.

**2] Profile Featurizer:** This is the second step and key component of Pipeline and It will convert raw data for each cluster into single numerical vector which can be used for machine learning algorithm in account scorer[2]. The extracted features are grouped into three categories Basic distribution features, Pattern features, Frequency features.

***a) Pattern Features****:* The authors have designed their own pattern encoding algorithm which maps the text to categorical variables and then take basic distribution features over these categorical variables. These features are used to detect malicious user especially bots which uses some pattern while sign up[2].

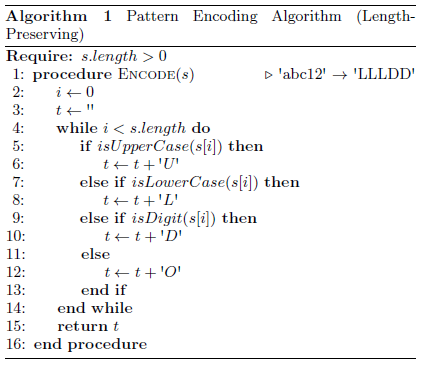


Figure 3: Pattern Encoding Algorithm which is Length Preserving[2]

The above algorithm is length preserving and will be able to detect 8 letters plus 3 digits in the email id but it won’t detect the list of email addresses of varying length so to overcome this problem it is recommended to use length independent variant algorithm.

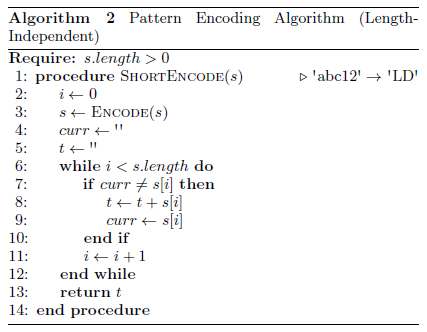


Figure 4: Pattern Encoding Algorithm which is length-independent[2]

***b) Frequency Features****:*  For each feature, calculate the frequency of the feature on the entire account database then compute basic distribution features on these frequencies. In general it is expected that the clusters of legitimate accounts may have either high frequency data or low frequency data where are as bots will show less variance in the frequency[2].

***c) Basic Distribution Feature****:* For each cluster, take basic statistical measures of each column. Examples include mean or quartiles for numerical features or number of unique values for text features.

**3] Account Scorer:** The account scorer takes as input the output of profile featurizer i.e., single numerical factor for each cluster. The specific learning algorithm that they have considered are logistic regression, random forests and support vector machines. These algorithms are used for dual modes, In “training mode” the account scorer is given a set of data and output the model description along with evaluation metrics which can be used for compare models. And in “Evaluation mode” the account scorer is given model description and the single vector value pf the model and output the score for the cluster indicating the likelihood of being composed of fake accounts. Based on the output score of the account scorer, the accounts of the cluster can be selected for three actions: Automatic restrictions, manual review or no action.

***a) Logistic Regression:*** Given a set S = {( x(i), y(i))}mi=1 of m training samples with x(i) as feature inputs and y(i) belongs to {0,1} then Logistic regression can be modeled as

D:\SPRING2016\Pictures\LR.PNG where is a set of model parameters[2].

Without regularization, logistic regression finds maximum likelihood criteria. With regularization, it controls to have fewer variables chosen in the model. In this paper they have used L1 penalization to regularize the logistic regression model[2].

The model parameters are computed using this

D:\SPRING2016\Pictures\L1Mo.PNG

***b) Support Vector Machine****:* The second learning algorithm is the support vector machine SVM[59,60,61,62]. The training dataset again consists of pairsD:\SPRING2016\Pictures\svm1.PNGD:\SPRING2016\Pictures\svm2.PNG. As it was recommended to use non-linear classifier, they used SVM with RBF( Radial basis function) in training. The RBF can be formulated as D:\SPRING2016\Pictures\svm3.PNG where r is called kernel bandwidth and is tuned based on results of cross validation.

***c) Random Forest****:* It is an approach that combines many weak classifiers to form a strong classifiers, for each decision tree, first sample with replacement from original training dataset in order to get new training set of same size. In each node of decision tree, chose m features randomly and split decision tree according to best possible split among m features. For a given new sample datasets the result model scores it by running sample through all trees and combination of the results in this problem of binary classification it is simply percentage of tree which gives positive result on sample[2].

**Summary of Experimental Data and Performance Analysis**

In this approach for training model they have used around 275,000 accounts collected from linkedIn which were registered over six month period and among these 55% of the accounts are declared as fake by the Security Team of LinkedIn. For training the model they used classification algorithms such as Random Forest, Logistic Regression and support vector machines. Moreover, In order to measure the performance of classifiers they calculated the Area Under Curve and recall at 95% precision[2]. Practically, the precision rates may be higher or lower depending upon the business needs but out of 3 classification algorithms, random forest was found to be efficient.

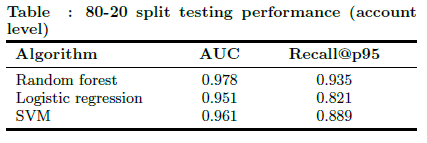


Table 5: Algorithms’ AUC and recall 95% precision[2]

The above table is data of all algorithms’ Area under curve and recall at 95% precision and the above experimental data is from the paper itself. And below is the ROC of all the algorithms.

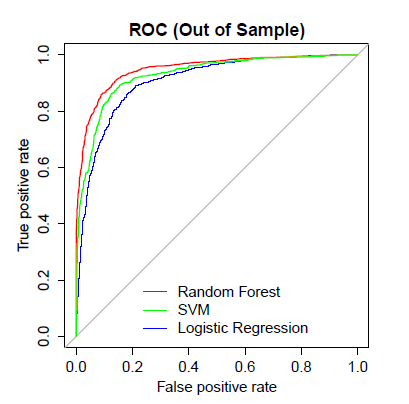


Figure 5: ROC of algorithms[2]

Below is the difference between out-of-sampling test performance between cluster level and account level.

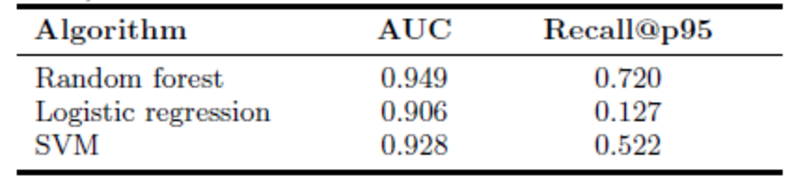


Table 6: Sample Testing Performance of Cluster Level[2]

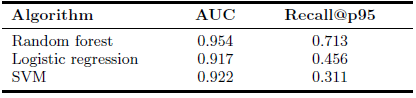


Table 7: Sample Testing Performance of Account Level[2]

The reason for random forests’ performance is better compared to other classification algorithms can be inferred from the table below .

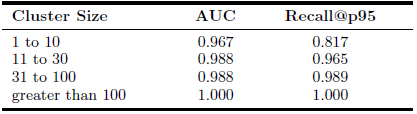


Table 8: Performance of Random Forest by cluster size[2]

**Strengths and Weakness**

1. As per the approach, the fake accounts are being detected at cluster level rather than account level so the time taken for detection is comparatively less.
2. The detection happens during the registration of account so this method is not allowing any fake account into OSN and it being discarded during the initial phase itself.

**Approach:**

The area of detecting fake accounts in Online Social Networks(OSN) is an increasingly innovative field, with an immense amount of research being poured into finding more efficient ways to identify fake accounts (Sybils).

While exploring the surface of this area of research, we came across several implementations and proposed systems that aim to reduce manual effort involved in scanning through each OSN account in order to determine whether it is malicious or legitimate.

Social networks in today’s world consist of a wide variety of interdisciplinary fields of study including graphs, social psychology, sociology and statistics. Due to the nature of the information shared by users of OSNs and the insight one gets into their personal lives, social networks are increasingly becoming attractive targets for malicious attackers and social engineers. We first approached the topic of detecting fake accounts in OSNs by doing an in-depth study on Facebook Immune System that was the first of it’s kind system to aid the detection of Sybils. We later move on to machine-learning based approach of detecting clusters of fake accounts in the entire OSN.

We further move ahead to a ranking-based system, SybilRank, that helps to rank each OSN user account based on certain confidence score. A listing of accounts is then the output of this system with the highly likely to be fake accounts at the top and the least likely ones at the bottom. This method takes into consideration that manual effort will definately be required in the process of detecting fake accounts and taking further action to deactivate or block them. But the system’s ranking makes it easier for humans in the process and helps them concentrate their time and effort into the most likely to be fake accounts.

**Coverage:**

Online Social Networks (OSNs) now-a-days have a huge user base and a lot of their personal information is exposed on the network as users rely heavily on the trustworthiness of OSN providers. However, OSNs easily become targets of fake account creations which can introduce spam, manipulate online ratings or exploit knowledge extracted from the network. OSN operators currently expend significant resources to detect, manually verify, and shut down fake accounts. Tuenti, the largest OSN in Spain, dedicates 14 full-time employees in that task alone, incurring a significant monetary cost [5].

In an attempt to automate the process of detecting fake accounts on OSNs, despite the difficulty in reliably capturing the diverse behavior of fake and real OSN profiles, SybilRank is a tool which relies on social graph properties to rank users according to their perceived likelihood of being fake (Sybils).

Based on a demonstration of SybilRank on Tuenti’s operation center it was found that ~90% of the 200K accounts that SybilRank designated as most likely to be fake, actually warranted suspension. On the other hand, with Tuenti’s current user-report based approach only ~5% of the inspected accounts turn out to be indeed fake [5].

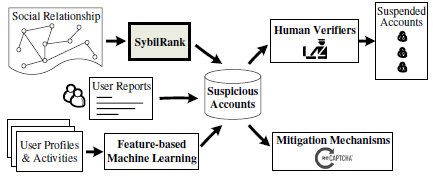
**What can fake accounts (Sybils) really do?**

Fake accounts can enable spammers in order to abuse the online messaging system of an OSN to introduce huge amounts of spam to the annoyance of the OSN’s users. They can waste an OSN by advertising customer’s resources by making him pay online, ad clicks or impressions to/ from fake profiles. Sybils may also gain access to users’ personal information on the OSN which can later be used in an adverse manner, e.g. social engineering to gain other useful information. Sybils may also manipulate online search results, perform large scale crawls over social networks [5].

**The challenge in detecting fake accounts using automations:**

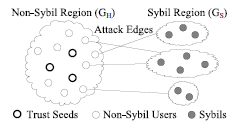
Due to the multitude of the reasons behind their creation, real OSN Sybils manifest numerous and diverse profile features and activity patterns. Thus, automated Sybil detection (e.g., Machine Learning-based) does not yield the desirable accuracy. As a result of this, adversaries can cheaply create fake accounts on OSNs and still manage to pass undetected. There is little evidence of wide industrial adoption of automated Sybil detection tools due to their shortcomings in terms of effectiveness and efficiency. Instead, OSNs employ time consuming manual account verification, driven by user reports on abusive accounts [5].

**How SybilRank works:**



**Figure 6:** SybilRank’s contribution in the OSN’s fake account detection process [5]

SybilRank is a Sybil inference scheme customized for OSNs that exhibit bi-directional social relationships. It is based on the assumption that OSN Sybils have a disproportionately small number of connections to non-Sybil users. The goal of SybilRank is to reduce as much human intervention in the process of detecting Sybils as possible. For this reason, it aims to output a quality ranking of nodes in the social network such that a substantial portion of Sybils ranks low. This helps to reduce the number of false positives or false negatives when it comes to Sybil detection.



**Figure 7:** Diagram depicting the Sybil and non-Sybil regions in the social graph

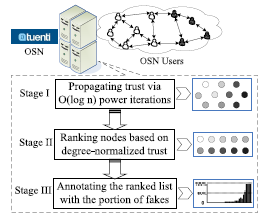
Let graph G contain ‘V’ nodes and ‘E’ undirected edges. Let Deg(v) be the degree of node ‘v’. It is assumed that the entire social network graph is divided into 2 disjoint sets of nodes- Non-Sybil nodes (GH) and the Sybil nodes (GS). The Non-Sybil region has very sparse connections to nodes in the Sybil region which are known as attack edges.

We assume that Sybils establish a limited number of attack edges due to the difficulty of soliciting and maintaining reciprocal social relationships with non-sybil nodes.The nodes in the Non-Sybil region of the graph are likely to be densely connected as opposed to the nodes in the Sybil region. So when traversing the graph from a trust seed (any randomly chosen Non-Sybil node) a sparse cut is encountered as the path is transitioned from the Non-Sybil to Sybil region.

SybilRank relies on the observation that an early-terminated random walk starting from a non-Sybil node in an OSN has a higher degree-normalized landing probability to land at a non-Sybil node than a Sybil node.

The key insight here is to rank nodes according to the degree-normalized probability of a short random walk that starts from a non-Sybil node to land on them. The low ranked nodes can then be screened out as potential fakes.

**SybilRank works in following 3 phases:**



**Figure 8:** Process followed by SybilRank in detecting Fake accounts [5]

1] Through w = O(log n) power iterations, trust flows from known non-Sybil nodes (trust seeds) and

spreads over the entire network with a bias towards the non-Sybil region.

2] In step 2, SybilRank ranks nodes based on their degree-normalized trust.

3] In the final stage, SybilRank assigns portions of fake nodes in the intervals of the ranked list. This

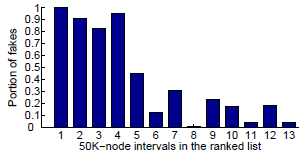
enables OSNs to focus their manual inspection efforts or to regulate the frequency with which they

send CAPTCHAs to suspected users [5].

**Experimental results & analysis:**

SybilRank was implemented on an experimental basis on Tuenti, one of Spain’s largest OSN with a huge user base and which provides multi-platform cloud experience to its users. It is also known as the “Spanish Facebook”. Prior to this experimental implementation, Tuenti employed 14 full-time employees especially dedicated to the task of detecting fake accounts from among millions of users and taking actions on accounts reported to be abusive. Apart from being a tedious and titanic task, a pure human view of the entire OSN is bound to contain errors.

SybilRank does not follow a priorly used machine learning(ML) based model, as ML- based systems tend to be too strict when it comes to the outcomes, and many a times may wrongly flag accounts as fake. They also have a significantly large number of false positives and false negatives. Based on Tuenti’s earlier user-based approach, ~5% of the suspected accounts would actually turn out to be fake. On the other hand, almost 90% of the accounts flagged by SybilRank as suspicious indeed did turn out to be fake. SybilRank also exhibited a ~20% lower rate of false positives and false negatives as compared to its contenders.



**Figure 9:** Distribution of accounts that are highly likely to be fake based on SybilRank’s score [5]

As described earlier, the output of SybilRank is a list with highly likely to least likely fake accounts. Starting at the lower end of this list, it was found that about 2K accounts that were flagged as suspicious were indeed fake. The fake accounts gradually reduce as we move up the list. In about 200K accounts in the lower end of this list, about ~90% of the accounts were truly fake.

SybilRank is a utility that researchers created as a method of ranking accounts based on if it is real or fake. In *[5]*, the authors describe their methodology and reasoning for creating this tool and measure its computational efficiency and accuracy at ranking accounts.

Using the OSN, Tuenti, the researchers found that of the 200,000 user accounts approximately 90% were fakes. Tuenti’s approach for detecting fake accounts was through users’ reports, however this approach only detected approximately 5% of fake accounts. The current approach by the OSN is rather expensive, since they have to employ 14 employees to only shut down fake accounts [6].

After running their application on Tuenti, the authors discovered different attributes of Sybil accounts and their relationships with other accounts. Since the application runs by forming edges between nodes, a tree structure would form between connected nodes of similar degree. This suggests that these Sybil accounts were created *en masse* for the purpose of perpetuating attacks on legitimate users. Furthermore, it is important to note that the Sybils do not form a single cluster but many smaller clusters. Therefore, this suggests that this is not a coordinated directive, but rather that there are many attackers who are creating Sybils [6].

**Strengths & weaknesses:**

The systems that we studied basically make an attempt to reduce human intervention required in the process of detecting fake accounts in OSNs. Based on the experimental results, these systems do seem to be effective in terms of pinpointing the accounts that are highly likely to be fake.

**Strengths of SybilRank for detection of fake accounts in OSNs:**

* It is a cost effective and robust system. The more the number of trust seeds chosen, the more robust and reliable the systems becomes. Increase in the number of trust seeds does NOT increase the implementation cost of the system.
* More effective than the manually driven approach which can be prone to errors.
* Saves manual effort, and also saves cost involved in employing humans to do the same task.
* Can be effectively used even for big OSNs with millions of users.

**Weaknesses encountered working with SybilRank:**

* Manual effort is still not completely avoided even with the use of SybilRank. Accounts that are ranked as highly likely to be fake need human intervention in order to decide what needs to be done further with them like presenting the user with a CAPTCHA challenge, asking the user to identify certain friends, or account deactivation. These decisions still need to be handled by humans.
* A fake account that has well established trusted links to non-Sybil accounts may still escape the detection by SybilRank, as the system relies on the assumption that Sybil nodes have very sparse connections to non-Sybil accounts.

**How do we deal with accounts that are NOT completely fake?**

**Approach:**

In new light, account creation may not just be legitimate or fake. Compromised accounts are being an increasing concern in OSNs too. A compromised account is a legitimate account that was maliciously taken over by an attacker. Since a normal account can have a multitude of interconnections with other users accounts on the OSN, the attackers tend to leverage this already established trust relationship while exploiting compromised accounts. We study the COMPA system which helps to detect such compromised accounts based on the accounts deviation from its normal, expected behavior.

The COMPA system leverages upon certain features of a user account which would characterize the normal, expected behavior of the person handling the account. For example the geographical location the user usually posts his messages from, the language he/she uses to share messages, the friends they frequently interact with, the URLs they share etc. The system then creates a behavioral profile of the users, and studies the deviation of new user posts from their behavioral profiles.

**Coverage:**

Fake accounts typically exhibit highly anomalous behavior, and hence, are relatively easy to detect. As a response, attackers have started to compromise and abuse legitimate accounts. Several attempts have been made in the area of detecting fake accounts (Sybils) on OSNs, however unfortunately such systems cannot discriminate between Sybils and compromised accounts. A compromised account is an existing, legitimate account that has been taken over by an attacker.

Compromising legitimate accounts is very effective, as attackers can leverage the trust relationships that the account owners have established in the past. Moreover, compromised accounts are more difficult to clean up because a social network provider cannot simply delete the corresponding profiles. Also, fake accounts can be safely deleted without affecting legitimate users. To address compromised accounts, however, the social network provider has to notify the victims, reset their passwords, and engage the users in a password recovery process.

In order to determine compromised accounts, this paper describes a way to model the regular activities of individual users. Hence, at any point if a user’s account gets compromised, there will be a noticeable change in the account’s behavior.

This approach detects similar shared messages that also deviate from the users’ normal behavioral profile. Off course we cannot detect cases where the attacker posts a single malicious message through a compromised account as the message might just be an outlier or merely reflect a normal change in the user’s behavior. Hence, this approach requires that a significant subset of these messages violate the behavioral profiles of their senders.

**Behavioral Profile of a user on an OSN:**

A behavioural profile of a user on an OSN is a collection of the expected behaviours of the user. The system leverages the user’s historical account activity data to build his behavioural profile. For example, the system could take into account all of the user’s posts, the photos shared, the people most interacted with, frequently accessed resources, the usual logging times, city, country of residence, time zone etc.

In order to create a user U’s behavioural profile, we first need to collect a minimum ‘S’ number of messages from the user’s profile, for e.g. the number of tweets on Twitter or the number of posts on Facebook. This will include the user’s posts on his/her own wall and also those on their friends’ walls. The messages collected are in a chronological order and the number ‘S’ is chosen to be sufficiently large so that enough messages can be processed about a user. If ‘S’ is quite small, the analysis may backfire and a legitimate user may be wrongly flagged as a compromised one. Certain newly created accounts or inactive accounts may not meet the ‘S’ number of posts criteria, but these can be ignored as they are less likely to be compromised. The system then extracts a set of feature values from each of the user messages in the input set of messages, and then for each feature trains a statistical model. Now given the behavioural profile of a user, we can assess to what extent a new message shared by the user corresponds to his/her expected behaviour. An anomaly score is computed for each new message that is posted by the user, based on which it can be decided if the new message has deviated from the user’s expected behavior. The anomaly score of a new message is computed by extracting the feature values of the message and then comparing them to the feature values of the user’s profile. The anomaly score is a value between 0 (perfectly normal) to 1 (highly anomalous).

**Feature extraction and modeling:**

Following are 7 features that are extracted from the input set of user’s historical messages and also while calculating the anomaly score of a new message:

**1] Time (Hour of the day):** This model captures the time of the day when the user’s account is most active (e.g. morning time, lunch time, evening) or when it typically tends to be dormant (e.g. night time). According to this feature, a new message posted as per the user’s active times is considered normal while those posted during dormant times might be malicious.

**2] Message Source:** This feature takes into account the application that was used to post the messages. The user may have used the OSN’s web application, mobile application or even a third party application to post their message. Off course, a third party application will require the user’s permission to post anything on their behalf. OAUTH is a well known software that is used by Facebook and Twitter to help users grant access to their profiles without sharing their credentials.

**3] Message Text (Language):** Although the user the flexibility to post messages in any language of his/her choice, there is a small set of languages the user is likely to use. These set of select languages may be considered stable for a particular user. So when the user posts messages in a different language, it may be an indication of suspicious activity with the account.

**4] Message Topic:** Users may post messages related to a wide variety of topics. But there are certain topics of specific interest to the user for e.g. music, artists, groups, companies etc. When users suddenly post something completely irrelevant and off track, it might be an indication that the account is compromised.

**5] Links in Messages:** Often messages posted on OSNs include URLs to other resources such as shared videos, blogs, pictures or new articles. In this case, analyzing whether the URL itself is malicious or not is not considered. Instead the fact whether the URL posted by a particular user is any different from his/her normal interests or not is of importance.

**6] Direct User Interaction:** This feature aims to capture a user’s interaction history with other users. A user may have a set of friends that can be directly interacted with via messenger or posts or tags. Analyzing this interaction history provides a good background on the people the user generally communicates with. An interaction that falls outside this set of friends, especially containing spam data could be a strong indication that the account has been compromised.

**7] Proximity:** In many cases, users tend to befriend people who are geographically close to them, for e.g. same city, same university, same country. If a user account is trying to interact with a person that seems to be on a different continent with no common ties, then this would qualify as a suspicious outlier.

The model now analyzes each new message based on the above mentioned features and computes an anomaly score value with respect to each feature. Finally each model’s score is combined to compute a consolidated anomaly score for the message. Based on this score, we can infer whether an account is legitimate or compromised.

**Experimental results & analysis:**

The proposed COMPA system conducted its experimental studies on two of the largest online social networks known, Facebook and Twitter. COMPA was tried on a dataset of about 1.4 billion user accounts in Twitter and about 106 million users on Facebook. COMPA successfully identified compromised on both the OSNs with high precision [4].

A large number of tweets for Twitter and posts for Facebook were captured and analyzed for this experimental study. COMPA was granted higher privileges on Twitter system and RESTful APIs to be able to capture live tweet streams from users, resulting in about 15 million tweets per day on an average [4]. The system then used an observation interval for one hour and scanned all tweets for the all the users at a particular hour. The system then used this data to construct behavioural profiles of users involved.

Earlier, creating honey accounts (e.g. profile with attractive pictures) would tend to attract malicious users, however this approach is not scalable. For Facebook, the dataset was obtained from an independent research group that conducted a study on Facebook. Unfortunately, Facebook actively prevents researchers from collecting newer datasets from their platforms by various means., including the threat of legal action. The dataset was crawled from geographic networks on Facebook covering areas of London, New York, Los Angeles, Monterey Bay, and Santa Barbara. The data collected spanned over two years of posts.

**Strengths & Weaknesses:**

The COMPA system is designed quite intelligently to filter compromised accounts from a huge dataset of modern OSNs. Detecting compromised accounts is a bigger challenge as compared to detecting fake accounts as compromised accounts are more deeply integrated within the social graph as compared to fake accounts.

The system tries to classify a user’s normal, expected behavior on the social network and an unexpected action or deviation from normal behavior. This means, any violation of the normal behavior will be recorded and if exploited repeatedly will classify the particular account as suspicious.

On the other hand, it might give rise to higher false positives and false negatives. For example, if a user moves to another time zone and posts his messages at a deviated time, or if the user learns a new language and posts messages in a different language than usual, it might still be captured as a suspicious account. Similarly, the malicious attacker might spread his/her messages at a perfectly ordinary time and in an acceptable language, hence might succeed in bypassing the feature extraction for compromised accounts and stay undetected.

**Studying the network and behavior of fake accounts**

**(i) User footprints in different OSNs**

Users possessing accounts across multiple social networks leave a digital footprint across the web that can be used to uniquely identify them by extracting data from public information on their profile.[6] These footprints can be collated to find out all of the user’s online identities, which has its benefits as well as security concerns.

**Approach:**

User Profile Disambiguation **-** Automatedclassifiersare trained with what are called similarity vectors, which are of the form < usernamescore, namescore, descriptionscore, locationscore, imagescore, connectionsscore >. These scores are formed from the similarity in data between a pair of accounts. Scores from account pairs which are already known to be of the same user, as well as of different users, are fed to the classifiers to train them to differentiate between same users and different ones. Four classifiers have been used in the model presented: Naive Bayes, kNN, Decision Tree and SVM. Once trained, this module gives a match/non-match result for account pairs like LinkedIn, Twitter etc.

User profile information across multiple services are collected through Google’s Social Graph API, and social aggregators such as FriendFeed and Profilactic.

Profile Features and methods used in making the digital footprint:

a) *UserID*: UserID is commonly the unique identifier that links a user to a social network. Since user ids of a user are not consistent across networks, “Jaro-Winkler distance”, a sophisticated method for string matching, was used to link the profiles, which gives a score in the [0, 1] range; higher value meaning higher similarity.

b) Display Name: First names and last names that a user has entered are matched with the “Jaro-Winkler distance” method. Clearly, this feature cannot alone be relied upon as the display name might be similar for many different entities.

c) Description fields: This refers to the ‘About me’ information that users provide themselves. Three methods were used for description field comparison from two profiles:

(1) tf-idf vector space model – Punctuation and stopwords are removed from the description and then the tokens are converted to lowercase in preprocessing. The cosine similarity between the two sets of tokens gives a score in the [0, 1] range.

(2) Jaccard’s Similarity – The preprocessing step is same as before, and Jaccard’s similarity gives a similarity score for the two token sets.

(3) ‘WordNet based Ontologies’ – Using the Wu-Palmer similarity metric between tokens of the description ﬁelds from the two user proﬁles to be matched.

d) Location: Four methods used in this feature - Sub Strings Score, Jaccard’s Score, Jaro-Winkler Score, Geographic Distance

e) Profile Image: Mean Square Error, Peak Signal-to-Noise Ratio, and Levenshtein were used to analyze profile image. They are first collected and scaled to 48 X 48 pixels and converted to grayscale. Each image is seen as a vector of values from 0 to 255.

f) Number of Connections – It was supposed that the number of friends that a user has across different networks would be similar. But in fact the number of connections could be very different and mean different things. Hence the connection values would first have to be normalized in the [0, 1] range for each service and then compared.

**Coverage and Experimental Data:**

Real world testing involved finding Twitter accounts using LinkedIn user display names, which resulted in 75% success rate when the top 3 results returned by the system presented matching accounts. 29,129 pair of accounts were collected from Twitter and LinkedIn for analysis. Accuracy, precision and recall were recorded at 98%, 99% and 96% respectively using the most promising set of features.

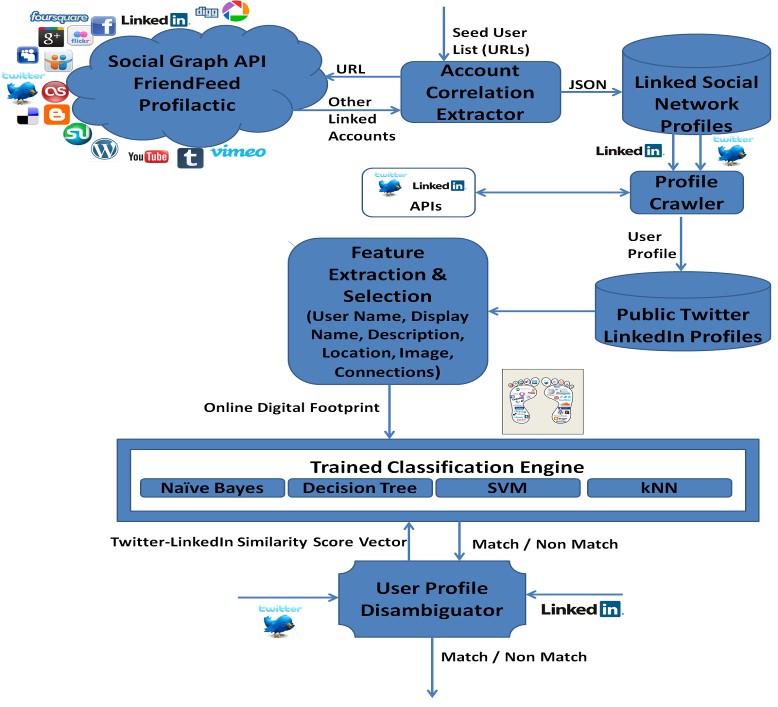


Figure 10. System Architecture showing data collection, feature extraction, training and results

**Strengths and Weaknesses**:

Among profile features, UserID, Name and Location are the most relevant ones as the experiments upon the model above suggests. Description and image came next, while relying number of connections was flawed due to the different nature of different services. If we have profile information that we are certain to be fake, then using the model we might be able to find fake accounts created across other networks by the same user. But more promising profile features need to be incorporated in the model for more accurate results. Also more social networks need to be included, currently the model has been tested only on LinkedIn – Twitter account pairs.

**(ii) Social Network Profile Cloning**

The method to be discussed relies on information contained in a user’s account that can uniquely identify them [10]. This ‘rare’ user information is used to find the user’s cloned profiles across other social networks. Any such accounts found are considered suspicious and more analysis is performed on them. The output is a list of similar profiles with a score attached to them to indicate the degree to which these profiles are similar to the original input profile.

**Approach:**

The approach consists of three components:

1. Information Distiller – This component takes a real social network profile and extracts the user-identifying information from it. It takes user profile information, runs test queries on the web, and those information that provide the least results are considered as the user-identifying ones. Along with the user’s full name, this information creates a user record which is passed to the Profile Hunter to identify potential social network user profiles that have matching information.
2. Profile Hunter – With the user record generated from the previous step, profiles that might potentially belong to the user are found from different social networks by using their specific search mechanisms. The hunter looks at the specific user information and the user’s real name, and from the results a profile record is created which contains the link to the user’s real profile and links to all the profiles returned in the results.
3. Profile Verifier – Each profile in the profile record is compared to the original profile in similarity of the values in information fields. Additionally, the profile pictures are also compared as fake profiles tend to use the actual user’s profile picture to look authentic. The comparison yields a similarity score for each profile which is presented to the user along with the list of all profiles.

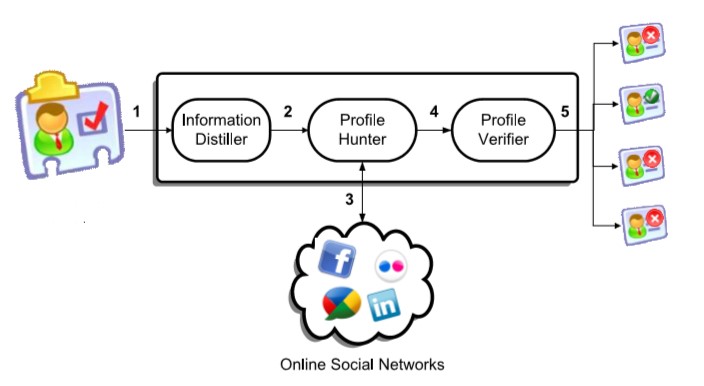


Figure 11. System architecture for detecting social network profile cloning

**Coverage and Experimental Data:**

Close to 30K LinkedIn profiles were collected by running queries on the traces shown in Table 9.

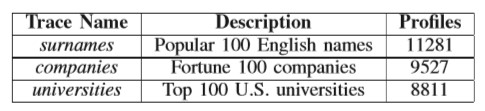


Table 9. No of profiles collected against the type of trace used

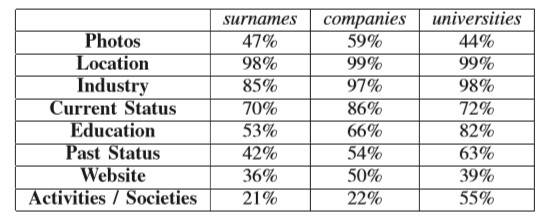


Table 10. For each trace, this table shows the LinkedIn public profile information

Table 10 shows the type of information available publicly in the different datasets from Table 9. In another experiment 1,120 public profiles were used (756 from surnames dataset, 224 from companies’ dataset and 140 from the universities dataset) to find duplicate profiles within LinkedIn. At least one clone was found for 7.5% of the profiles. Only exact field matching was used for the experiment, profile pictures were not considered.

**Strengths and Weaknesses:**

The model only uses the LinkedIn network, which is a crucial drawback. It needs to be expanded to work on other social networks as well. While doing so, extraction mechanisms specific to those networks should be developed, and a single representation format should be designed for comparison. Also the reliance on exact word matching is flawed when information could be typed wrongly or changed deliberately by attackers. Hence alternate matching techniques such as fuzzy string matching have to be employed to efficiently find duplicate profiles.

**(iii) Identity Clone Attacks - Multiple-Faked Identities Profile Similarity (MFIPS)**

Instead of creating a single fake identity of the victim, an attacker can create multiple fake identities as well that are related to the victim [11]. This helps the attacker to forge the victim more easily.

**Approach:** The detection process on OSNs is based on profile similarity measures. The process has different phases • Discovery: An input identity (IID) is taken and compared against a set of profiles for similar ‘Name’ and the results are put into a Candidate List (CL) • Compute Profile Similarity: Each entry in the CL is compared with the IID using the profile similarity schemes, and the a predefined threshold is defined, above which the a similarity causes the CL to be put into a Suspicious Identity List (SIL). Since we cannot initially assume that the IID itself is authentic, it is also added to the list •Validation: Every entry in the SIL is validated, and the fake ones identified are put into a Faked Identity List (FIL). If an entry is genuine, its trust value is increased to avoid future validations. The fake ones are closed or deleted and the friends of those identities receive notifications of the same.

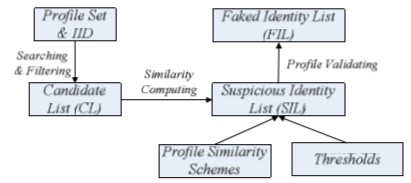


Figure 12: Detection process for Identity Clone attacks

One approach to validate an identity is to ask the user to provide a real world ID that would be unique, e.g. passport, but this may not be generally applied to all real world users. Moreover, if the attacker knows this ID somehow, then he will pass the identity validation test. Another approach, proposed by ‘mysafeFriend’, a third-party application in Facebook, is to have levels of trust. At lower levels of trust (1 to 3), the identity is verified by its friends. When enough friends identify positively, the trust level gets a promotion. At levels 4 and 5 information such as credit card detail is used to verify the identity, but such private information may not be comfortable for many users to divulge.

A better approach might have been to get an identity’s friends to set questions that only a real identity would have been able to answer, and then consider the percentage of correctly answered questions to ascertain validity. This is known as Social authentication. It is considered to be an ideal approach as it does not involve sharing of any private information. Another approach is to keep monitoring a suspicious account and analyse its activities, but practically it is time consuming and not guaranteed to give results.

**Coverage and Experimental Data:**

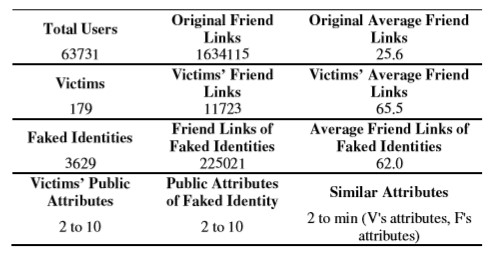
63,731 Facebook users have been considered for testing with a total of 1,643,115 friend links. 

Table 11: The table shows the experimental results.

**Strengths and Weaknesses:**

The model has a few drawbacks as follows: • When the victims themselves do not have a social network profile or may not even exist, the model is unable to detect such an attack. The friends who accept such an identity may become the victims in the future • An adversary can first create a genuine looking account that never used social networking (or does not even exist), earn enough trust and then change the attributes of the profile to victimize another identity. In such a case, the detection system has to be run when the profile got created or when the profile identifying attributes were changed so that they could have been detected • Real OSN testing has not been done for the model yet, one problem being that creating fake account for testing purpose in real world network will have damaging effects.

**(iv) Clustering of fake accounts in OSNs**

**Approach:**

This model has already been discussed in the report. The focus of this method is to identify clusters of fake accounts that are registered by the same user [2]. A supervised machine learning pipeline technique is proposed to achieve the required result, which is to classify a cluster of accounts as fake or real. This model primarily focuses on user text inputs to find patterns across multiple accounts in a cluster.

The *need for cluster level grouping of accounts* is justified with the following challenges in current fake account detection methods:

(a) Scalability: There exists algorithms to predict fake accounts at an individual account level, but as the OSN grows to a very large scale, these methods are too slow to effectively stop fake account creators from achieving their malicious purposes,

(b) Time-sensitivity: It is more desirable to stop a fake account as soon as it is created, but current mechanisms rely on user activity to classify them appropriately, hence a method to detect fake account patterns using the minimum account information is more desirable.

**How these challenges are solved**: 1) The model takes clusters as inputs for learning rather than individual accounts 2) The model only uses profile data from registration time to detect fake ones early.

**Coverage and Experimental Data**: It was observed that the model was capable of finding 250,000 fake LinkedIn accounts from 15,000 clusters. At one point, the precision dropped because the model was incapable of differentiating a large number of organizational signups from fake ones.

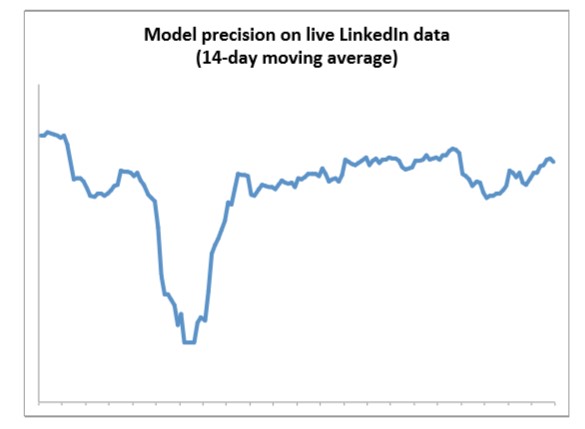
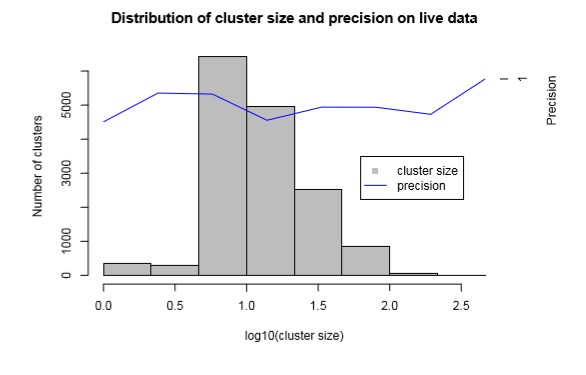


Figure: 13 Distribution of cluster and precision on live data(left) and model precision on live LinkedIn data(right)

**Strength and Weaknesses**: The method provided some false results such as grouping accounts originating from the same organization as fake. In that particular case, an alternate account detection model had to be incorporated to consider these accounts as ‘manual review’ category always. This involves human intervention which we would like to avoid. Also smaller cluster sizes might present some false positives, because the model relies on account patterns becoming diverse when put in a large cluster. Since cluster size is configurable, chances are present for this scenario to happen.

**(v) Case Study, which shows the privacy leakage of normal users to the fake account users**

**Approach**

In this case study [8] we see how there is a privacy leakage in Pengyou, one of the popular social networking site of China. Eight fake users were created as a part of the study. They gained a lot of online friends by sending friend requests to unfamiliar people and got the privilege in accessing the cyber-friends’ profile. Based on the eight fake account users and their collected online users’ profiles we analyze the degree of online self disclosure, the age distribution of fake account users’ friends and photographs leakage of fake users’ friends.

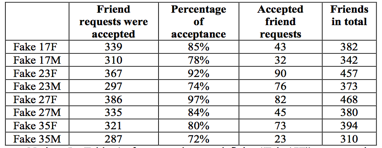


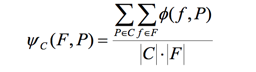
Table 12: overview of collected data [8]. Each fake user sends 400 requests.

**Coverage and summary of experimental data**

**The degree of online self-disclosure(DOSD):**

Social networking users post their personal information such as pictures, hobbies etc. Field leakage and degree of online self-disclosure is used to quantify user’s tendency to disclose his/her personal information.

Given a set of field names F , a user’s profile P , and a class of users C , we define the average DOSD of a class of users as:



Where, |C| represents the size of a class of users C .

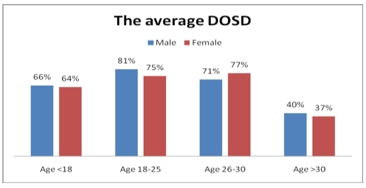


Figure 14 :The average DOSD for the different classes of users are shown [8]

**Photograph leakage:**

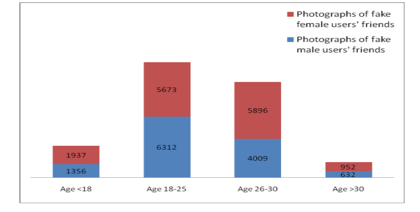
By creating the above eight fake users with 2671 friends in total, they succeeded in gaining access to a significant number of these users’ personal photos. 

Figure 15 : collected photographs data from these fake users’ friends[8]

**Threats:**

*Advertising:* By getting the access to the various personal information like personal interests, Email, occupation etc., several advertising companies takes advantage of this information and even can result in spamming which has become a major problem.

*Cyberbullying:* Users with fake accounts can publish news, reviews and other multimedia material to defame or attack people who may not be aware of the attack.

Organization’s reputation: Disclosed information in social networks can tarnish organization’s reputation and may also invite other avenues of attacks by cybercriminals and the advanced persistent threats can cause severe financial damage.

**Mitigating the Risks:**

Personal Privacy Settings: Social networking site should provide the options to hide personal sensitive information, prevent user’s friends annotating the profile with clear meaning words and preventing specific people access to user’s annotations.

Owner’s confirmation: Every operation which involves user information should be confirmed by information owners.

Developing the training curriculum: Before creating a new user account, proper demo or animation should be given about the risks of privacy leakage and how to set the privacy protection.

**(vi) Multiple Identity/Fake Account detection using non-verbal behavior**

We study the representation of nonverbal behavior for fake account users/identity deception or sockpuppets [7] , which can be applied in many types of social networks. Wikipedia is used as an experimental case.

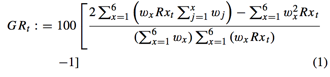
**Approach:**

Using the variables that represent non-verbal user behavior using Wikipedia is analyzed. Based on the behavior of Multiple identity users from normal users over the period of time, a detection method is proposed. The method is used on the testing data to analyze the accuracy and computational efficiency.

**Coverage:**

Wikipedia is a free online encyclopedia where everyone can contribute with an account using a pseudonym or real name. It operates on concept of namespaces with total of 28, where each namespace meant to include specific type of content. Wikipedia account users leave revision footprints on pages whenever they make changes to them. Wikipedia uses a page revision log maintained for every page in which everyone can find who did a revision, what is that revision and other things like when it was made, how many bytes are added or removed from that page. This logged data on page revisions will provide non-verbal user behavior. For example, time taken between each revision is considered as measurable non verbal behavior.

Simple and complex variables are used to represent user behavior. These variables of non-verbal behavior falls under two catagories like time dependent (variables are represented with index t) and time independent. The following are the variables considered. Rt - Number of total revisions made by a user for a specific time window since their registration with the website. The number of revisions is distributed in various namespaces like article (Rat), user page (Rut), article discussion (Rdt), user discussion page (Rtt). Wikipedia-related pages and related discussion pages combined under variable (Rwt). The other variable Rot is assigned for namespaces such as file uploads, images etc. Gini coefficient is also used that represents differences in activity distribution across the above namespaces which bounds between values 0 and 1.



where x represents all the revisions on each namespace mentioned above and w is the relevant weight assigned to each item which is equal in this case. Numbers of bytes added (Bat) and removed (Brt) during observation window are also measured. The time difference in seconds from first revision since user registered (TR) along with the namespace (FE) where their first revision was made. The average duration (ADt) is measured.

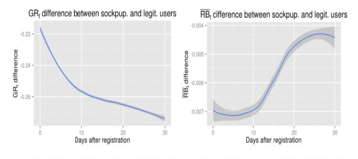
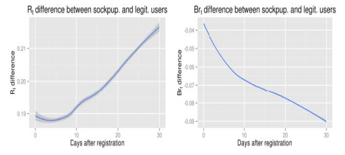


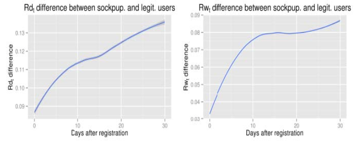
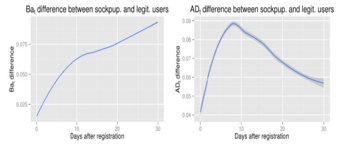
where n is the total number of revisions and T is the set of all Unix times for each revision made.

**Summary of experimental Data**

List of publicly available logs of blocked users on Wikipedia during the period from February 2004 to October 2013 are collected. These logs include various reasons for blocking user accounts for verified sockpuppet cases. About 38.96 percent of sockpuppets have their accounts blocked during first revision on Wikipedia i.e. on the first day. It raised to 62.24 after 10 days. By 30 days, the percentage rises to 74.43.

Final user list contained 7,500 verified sockpuppet cases and 7,500 legitimate user cases. For each one of the users in the above sample, all activity on Wikipedia is obtained. This activity can be translated into variables which mentioned above.





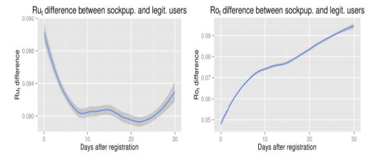
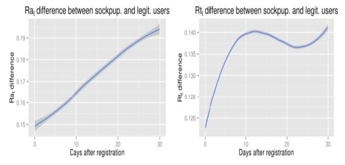


Figure 16 : The above figures are the variation of differences in non-verbal user activity variables between sockpuppets and legitimate users over a period of a month. Positive scores along the y axis indicate increased activity for sockpuppets whereas negative scores indicate decreased activity for sockpuppets compared to legitimate users[7].

**Accuracy:** A sample of 12,723 users are taken in which 48.23% of users are sockpuppets. Using the proposed method, 71.3% accuracy is obtained and the computational efficiency is O(I\*R’), R’ is limited amount of revisions(R’<R) made by user in window of observation.

**Strengths and Weaknesses:**

· Proposed method is viable and efficient method for detecting identity deception. It incurs a lower computational overhead over other traditional models.

· On the flip side, the efficiency and effectiveness of the method is influenced by several specific factors. If computational efficiency is not taken into account, the accuracy using the method is less compared to some machine learning models like SVM.

**(vii) Anomalous User Behavior in Online Social Networks using PCA**

Increased crowdsourced information led to a black market promotion techniques using fake and compromised accounts / collusion networks. There are supervised and semi supervised approaches for detecting these anomalous users. This study represents the technique called Principal Component Analysis (PCA) that distinguishes the behavior of normal users from anomalous users by identifying significant deviations [9].

**Approach:**

In the first step of PCA based anomalous detection, the user behavior is captured in a small number of dimensions. In the following analysis, the two years of complete user behavior data is used from nearly 14K Facebook users, 100K Twitter Users and 92K Yelp users. The behavior of normal users in these social networks is captured in the top 3 to 5 principal components. The user behavior, which cannot be captured by these principal components, is said to be anomalous behavior. 2nd Step is the evaluation of accuracy of the above PCA-based anomaly detection on ground-truth data taken from set of normal users and anomalous users from Facebook.

**Coverage:**

PCA is used to identify dimensions or features that explain predominant normal user behavior at its best. It projects high-dimensional data into a low-dimensional/normal subspace of top N principal components which accounts for the data with much variability. The projection onto the residual subspace/ remaining components detects the anomalies and noise present in the data.

The input features to PCA for capturing user behavior are as follows:

Temporal Features: These are time-series of values observed. The value here is number of likes per day granularity. The value may be time series of comments, number of posts, chat messages. The principal components from the input set can model rate of change in like activity, weekday and weekend patterns, inter like delay and other latent features.

Spatial Features: It is defined as histogram buckets of observed values. The category of Facebook pages are used as buckets and the number of likes in those categories is taken as observed value.

Spatio-Temporal Features: It combines the above two features which identifies evolution of spatial distribution of observed values . Entropy is used to summarize the distribution of the categories which are liked.

**Summary of experimental data:**

As mentioned above, the two years of complete user behavior data is used from nearly 14K Facebook users, 100K Twitter Users and 92K Yelp users. Capturing the normal user behavior for this data using the presented PCA analysis is shown below.

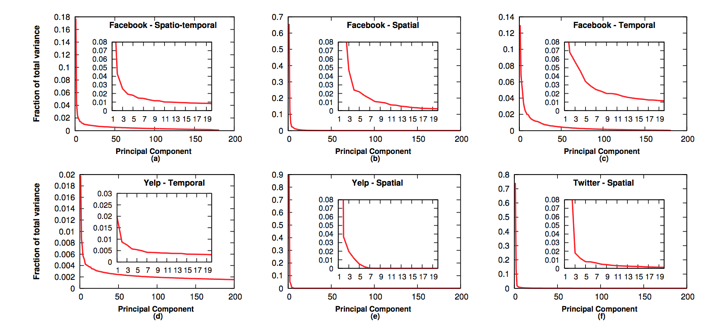


Figure 17 : Scree plots showing low-dimensionality of normal user behavior. A significant part of variations can be captured using the top three to five principal components (the “knee” of the curves). [9]

**Detecting anomalous User behavior:**

Training Data consisted of 12k random users sampled from Facebook. Testing data consisted of 3.2k users of Black-market, 1k compromised accounts, 900 colluding users and 1.2k normal users. After testing the results are as follows: 99% of the Black-market users , 64% of compromised accounts and 92% of colluding accounts are marked as liked flags. When tested on normal users, 3.3% are marked as false positives.

**Strengths and Weaknesses:**

· Principal Component Analysis provides more systematic and general type of framework for modeling user behaviors in online social network rather than the other techniques which are ad hoc and complex.

· On the flip side, PCA analysis may not be ideal for small crowdsourcing systems and they are not attack specific.

**What attributes lead to the success of fake accounts?**

**Approach**

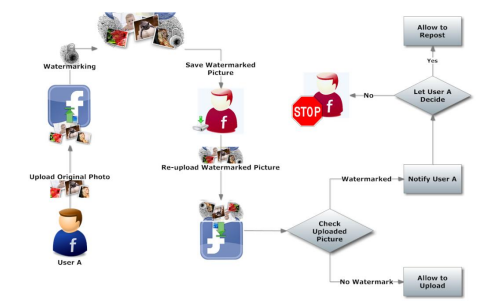
Fake accounts on OSNs is one of the sources of online attacks. Few may be harmless like the ones created to get more number of likes but others can cause serious damage like influence trending topics, influence political content, spread spam advertisement, tricking users to install malware and extract sensitive information, etc. The security vulnerabilities in these OSNs are responsible for success of fake accounts and by studying these attributes that aid fake accounts to succeed, countermeasures to attacks on OSNs can be developed. These attributes can be classified into two categories. One category focuses on approaches used to successfully create fake accounts and the other focuses on the approaches that aid fake accounts to exploit OSNs for their benefit like social engineering. We discussed security measures[1] employed by few popular OSNs and the mechanisms used by fake accounts to breach them.

**Coverage**

The following attributes contribute to the success of fake accounts in OSNs according to literature[3,13,14,15].

**(a) Fake email addresses** – There are services available which provide temporary email addresses (10minutemail.com). An attacker uses these temporary addresses which are valid for a few minutes which is enough to trick the OSN to accept the email address and creates fake user accounts. The adversary uses these fake accounts to connect with random or targeted OSN users and establishes fame and then profits from the data collected from victims. Often OSNs use two step verification to validate a user by sending verification code through email or a text to a valid email or phone. An attacker can use websites like ailhazard.com and numberproxy.com to generate dynamic email ids and phone numbers which can be used to bypass email/text/call verification tricking OSNs [3].

**(b) Fake User profiles** – An adversary employs social engineering techniques to attract users. By studying the attributes that are shown to be effective in getting users’ attention like using a picture of attractive man or woman as profile picture, these malicious profiles attract many users. An adversary grows its network by connecting with targeted or random people and misuse legitimate account user’s trust or information for their benefit thus abusing OSN. Facebook has taken countermeasures by watermarking every picture to prevent attackers from using others photos, but the attacker still uses photos by modifying them. As shown in the figure below, whenever a user tries to upload a picture to Facebook, it checks whether its watermarked or not. If the picture belongs to another user, the owner gets notified. But this might not work when the user trying to upload picture crops some part of the original picture or change it [3,13,14].



**Fig. 18 Proposed model to detect cloning [14]**

**(c) Automatic Account creation tools** – Automation by scripts or tools is the most economical way to get through the security challenges or create fake accounts [14,15]. The tools can be chosen based on the functionalities they offer and then configured to meet requirements.

Below are some of the popular tools used to create fake accounts and handle challenges.

**1. Twitter Account Creator Bot**

**Tool:** Twitter Account Creator Bot 2.0.0.6

**Source link:** http://sourceforge.net/projects/kipesoft-acb/

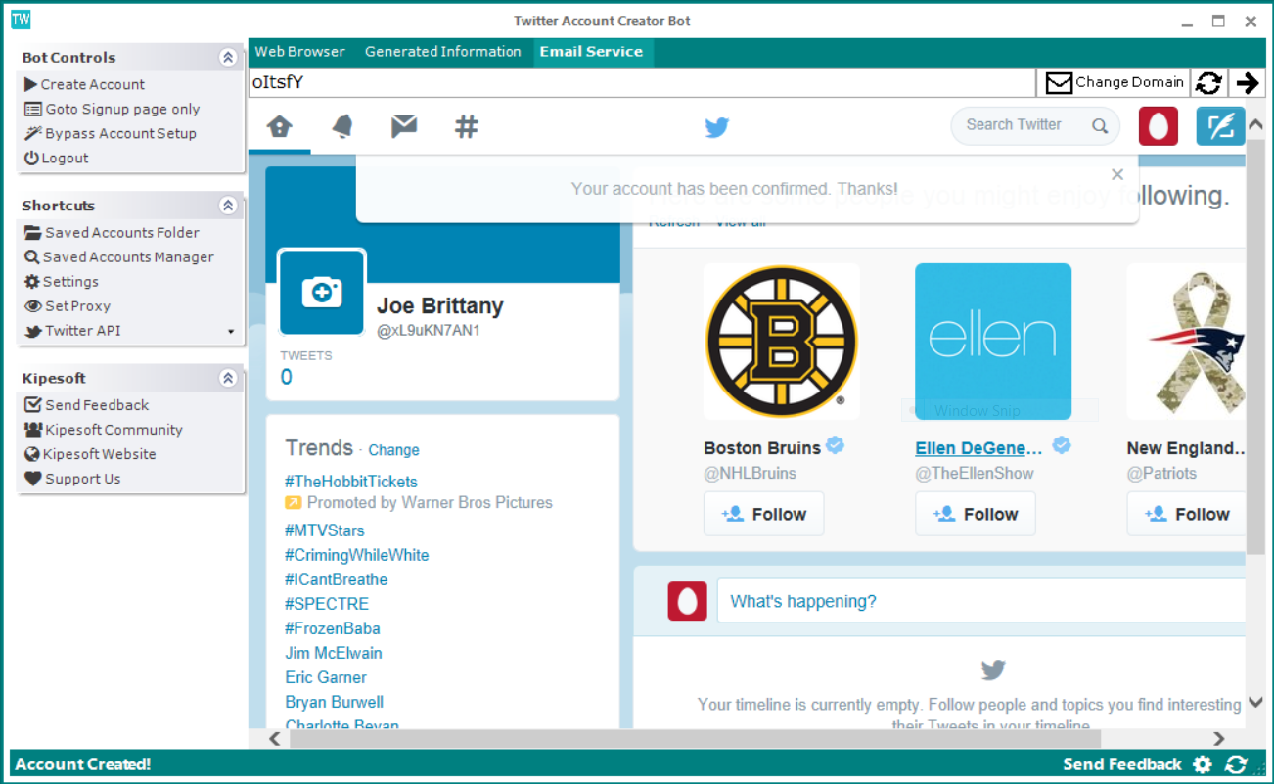
**Active Since:** 10/18/2014

**Platforms:** Windows, Linux, Mac

**Cost:** Free / Open Source

**Features:** Automated Following

**Challenge Bypass:** CAPTCHA Input, Email Verification, Proxy Support



**F**i**g. 19 Twitter Account Bot Creator [15]**

**2. FB Mass Account Generator**

**Tool:** FB Mass Account Generator, 4.0.0

**Source link:** http://www.latestautomationbots.com/facebook-mass-account-creator/

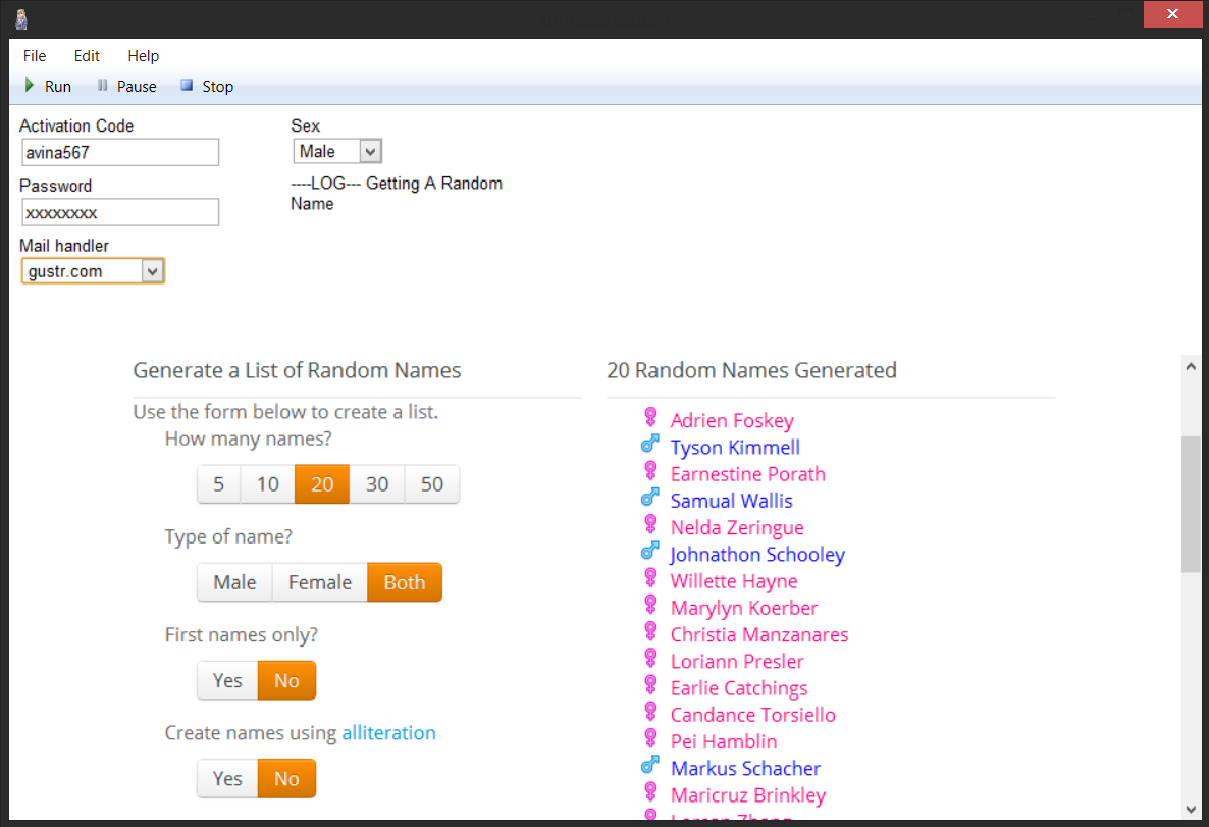
**Active Since:** 5/6/2014

**Platforms:** Windows

**Cost:** $7 per month

**Features:** Automated Liking, Following, Unfollowing, and Commenting

**Challenge Bypass:** Email Verification



**F**i**g. 20 FB Mass Account Generator [15]**

**3. PinMass**

**Tool:** PinMass 4.0.0

**Source link:** http://www.latestautomationbots.com/pinterest-mass-account-creator

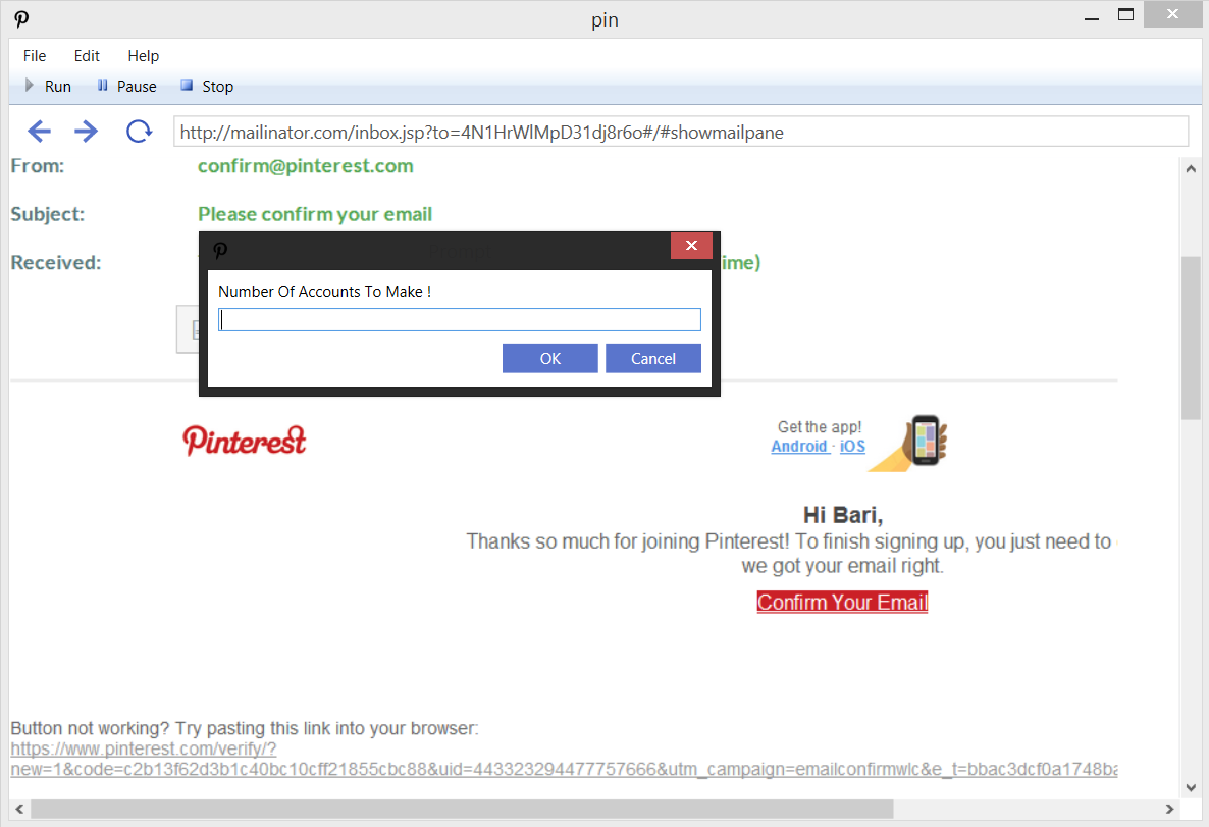
**Created on:** 5/1/2014

**Platforms:** Windows

**Cost:** $7 per month

**Features:** Automatic Following and Unfollowing

**Challenge Bypass:** Email Verification



**F**i**g. 21 PinMass Bot [15]**

**4. FA Creator**

**Tool:** FACreator 1.0

**Source link:** http://www.blackhatworld.com/blackhat-seo/black-hat-seo-tools/

663405-get-facreator-fast-web-2-0-account-creator-100-supported-websites.

html

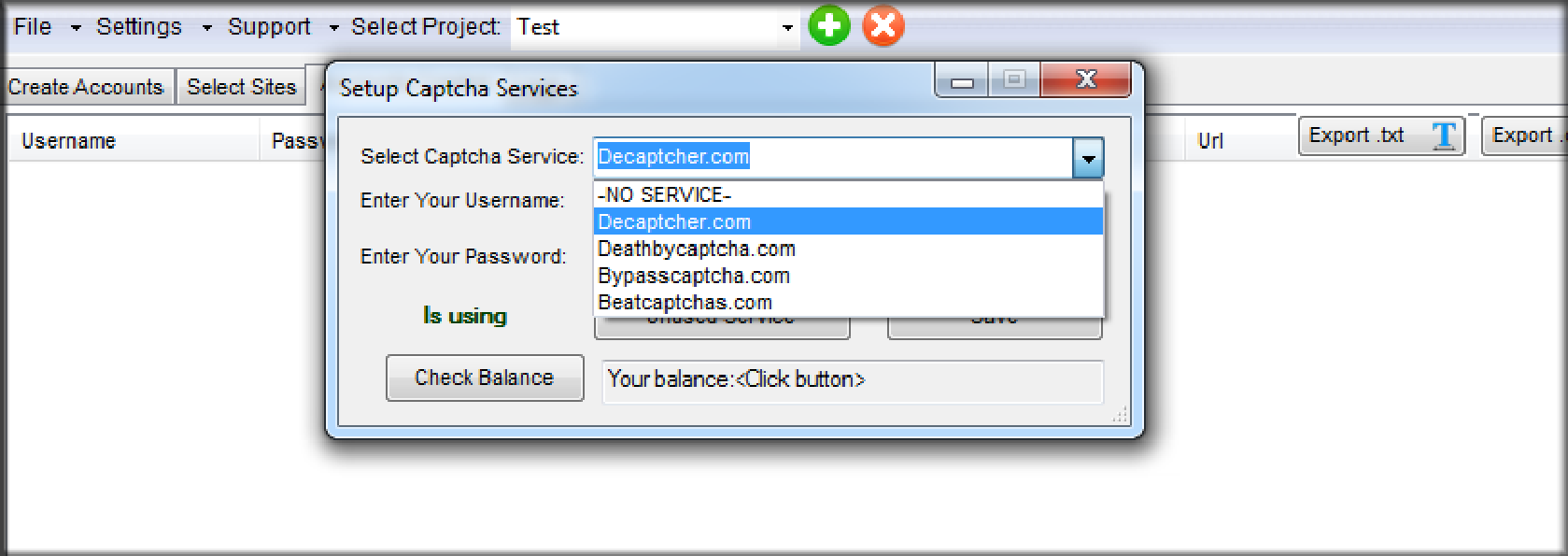
**Active Since:** 4/3/2014

**Platforms:** Windows

**Cost:** $12 per month

**Features:** Creates Accounts on Many Websites, Automated Content Posting

**Challenge Bypass:** Captcha Resolution, Proxy Support, Security Question Bypass, Email Verification



**F**i**g. 22 FA Creator [15]**

**5. Account Creator Extreme**

**Tool:** Account Creator Extreme 4.2

**Source link:** https://www.blackhatspot.com/Thread-GET-Account-Creator-Extreme-

4-2-25-Supported-Websites

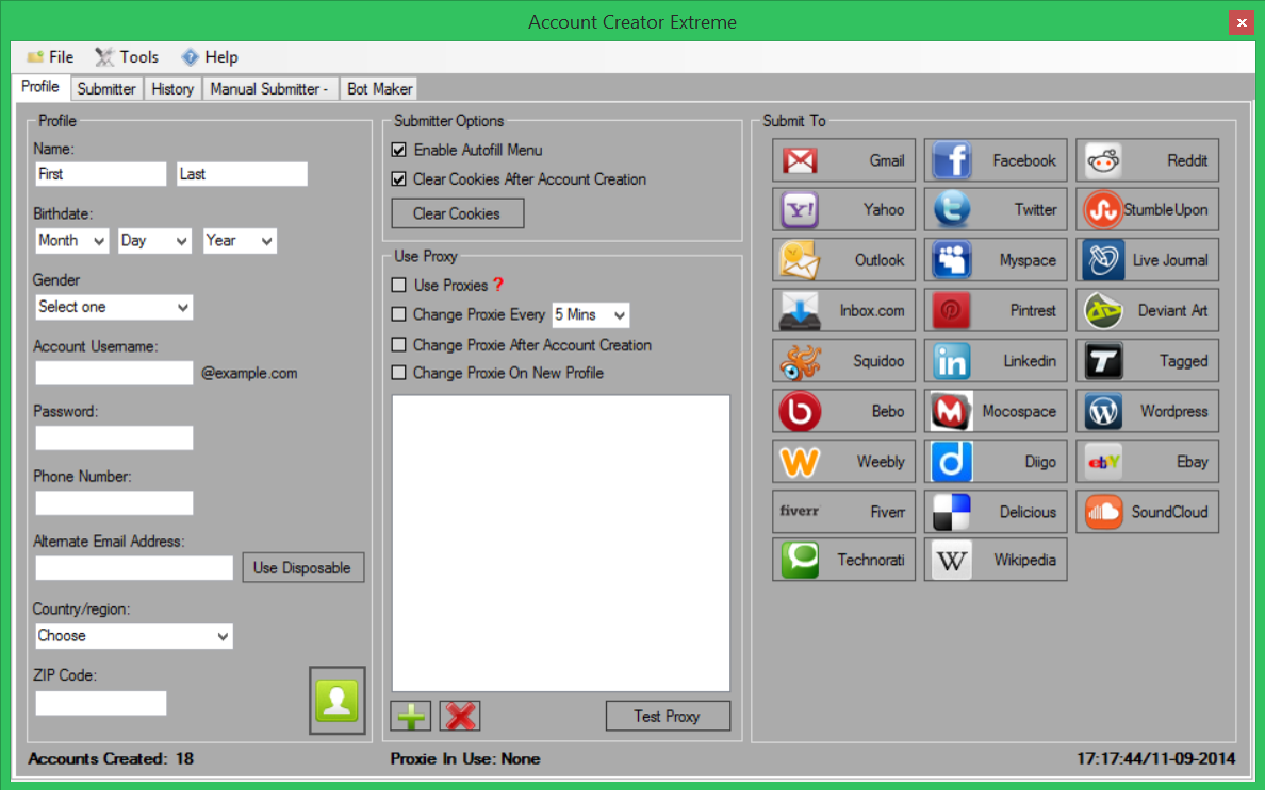
**Active Since:** 3/17/2011

**Platforms:** Windows

**Cost:** Free

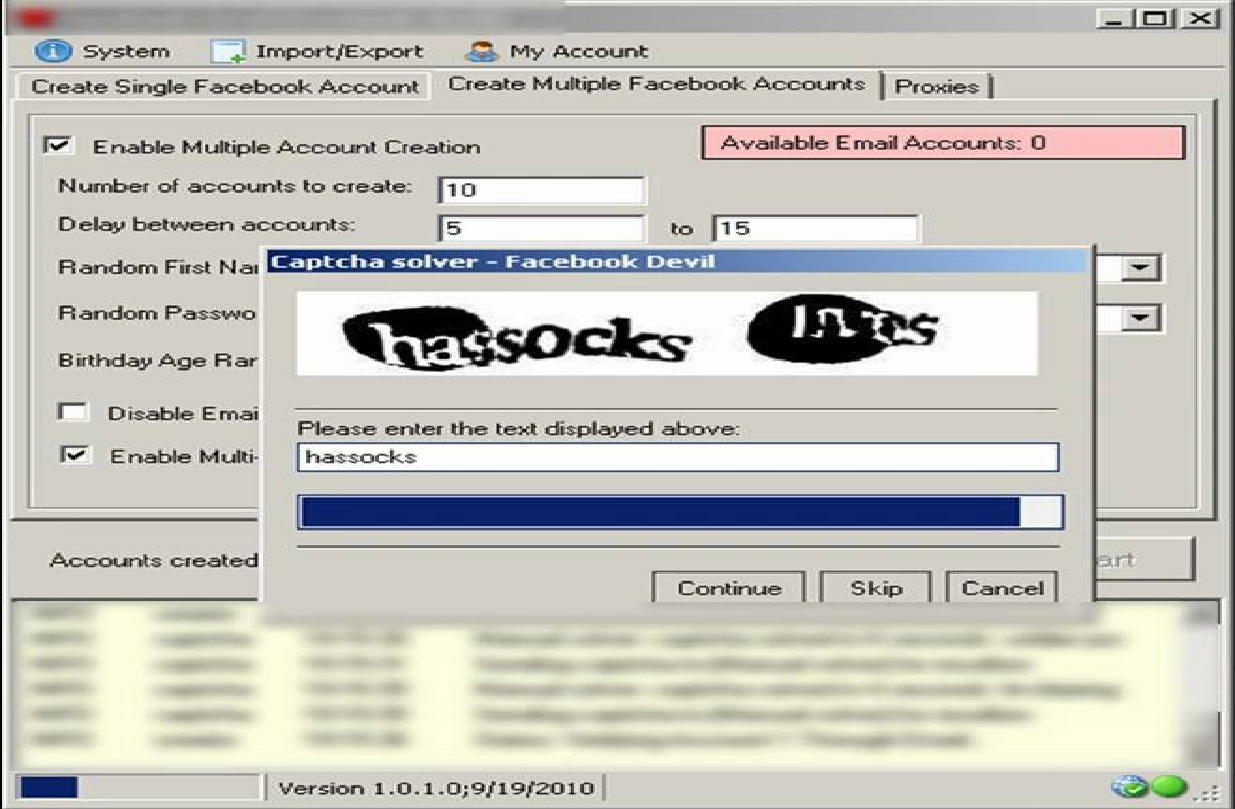
**Features:** Creates Accounts on Many Websites, Bot maker

**Challenge Bypass:** Proxy Support, Captcha Resolution



**F**i**g. 23 Account Creator Extreme [15]**

**6. Facebook Devil**



**F**i**g. 24 Facebook Devil tool [15]**

**(d) Knowledge of detection tools** – Attacker who knows how a detection tool works can take related measures to avoid detection of the fake accounts. According to the Literature[3,5], the knowledge of matching behavioural profiles and similarity measures can help to evade the detection tools.

**(e) Ineffective CAPTCHAs** - Socialbots are computer softwares which mimic real users and establish trust with other users thereby grows its network. These socialbots infiltrates the OSN and collects user’s private information and email addresses and the adversary profits from this data collected thus violating victims’ privacy and abusing OSNs. CAPTCHAs are used by OSNs to prevent automated Socialbots from misusing OSNs. However, CAPTCHAs can be broken by tricking a human victim to solve CAPTCHAs or by hiring cheap labour (there are CAPTCHA- breaking businesses which pay users for breaking CAPTCHAs). Some websites challenge users with modified CAPTCHA like solving a math puzzle or identifying an image. But the bots store these challenges in database and update new challenges [3, 13].

**(f) Exploitable APIs and Platforms** – To make integration of an OSN platform into different third-party software systems easier, most OSNs provide software APIs. If there are any vulnerabilities with the platform of OSN or within any API used by the platform or provided by OSN, an adversary can exploit these vulnerabilities and can abuse the OSN. For example, Facebook provides Graph API which enables third party softwares to read and write data to Facebook and provides a simple view of social graph with representing objects (profiles or photos) and their connections with other objects (friends or tags). An adversary can use this API to perform online social activities. OSNs are found to be affected by XSS worms [3].

**(g) Crawlable social graphs** – OSNs usually hide the social graphs to protect the privacy of its users. However, an adversary can generate parts of this graph by creating a fake profile and crawling through the profiles connected to the initial ‘seed’ profile like a Facebook user can go through their friends’ friend list and can generate the social graph. There are some commercial crawling services supporting OSN crawling like 80legs.com and can effectively generate social graphs. An attacker can either use these crawling services which are inexpensive or build customized web crawlers for this task [3, 13].

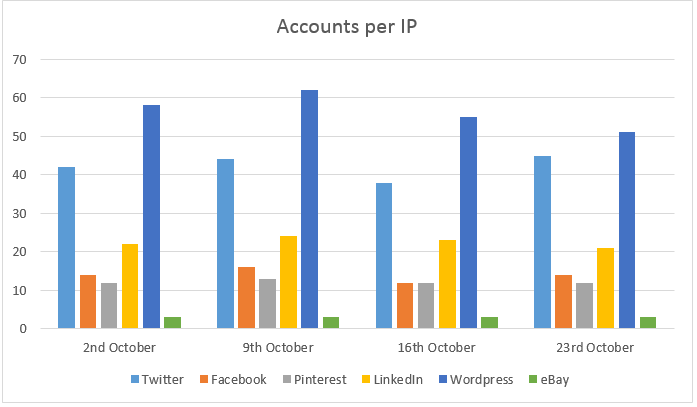
**No effective security and privacy controls** – There are no effective privacy and security controls that OSNs can employ to protect or warn users against befriending malicious users. Designing controls that better communicate the risks of befriending a stranger or an attacker require analysing and eliciting the befriending behaviour of users and the factors that influence their decisions.[13]

**Data & Analysis**

Three factors are being considered to analyse popular websites according to literature [14].

* Number of accounts that can be created from a single IP

Ebay has restricted number accounts created for a single IP to three. Facebook, LinkedIn & Pinterest allowed creation of 20 accounts without any challenges from a single IP while Twitter allowed > 40. Wordpress on the other hand has no security measures implemented to halt the fake account creation. Below is the statistics showing the account creation testing process on various websites.

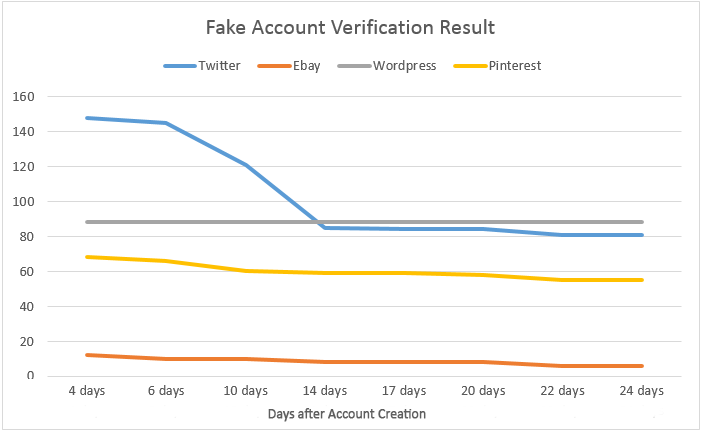


**Fig. 25 Experimental results showing the number of accounts we could create on various websites using a single IP address[15].**

* If the fake accounts created were identified and banned

OSNs should implement a mechanism to detect existing fake accounts and suspend them. Out of all the websites under study only Twitter has an observation period of 6-10 days before identifying the set of IPs created from a single IP and marking them as suspicious accounts. Some websites mark fak accounts based on their activity or reporting by other users.

Below is the graph showing the Account verification activity for the fake accounts created by the testing team



**Fig. 26 Experimental results showing how many of our created accounts were banned or suspended by various websites[15]**

* Effectiveness of counter security measures implemented by the websites under study

*Twitter* - In the above experiments, Twitter Account Creator bot was used to create fake accounts. For every new account, an email verification is required. But when 40 accounts are created from a single IP, Twitter presented a phone verification challenge. As this tool does not support this feature, the account creation is stopped. But this always be bypassed by changing IP or choosing a tool that has phone verification feature.

*Facebook* - It has a email verification for every account, after 14 accounts, presented a CAPTCHA, after 5 CAPTCHAS asked for a phone verification. The attacker could still continue by changing his IP.

*Pinterest* - This has only Email verification.

*LinkedIn* - It presents with CAPTCHA after every 5 accounts from same IP. Just like others, this can bypassed by switching IP using DeCaptcha to solve CAPTCHA.

*Ebay* - As Ebay is a marketplace, strict measures are enforced for account creation. Three accounts per IP per day can be created making it slow for mass account creation.

*Wordpress* - No measures implemented.

**Strengths and Weaknesses**

The papers referred discuss various well known security measures in practice and in theory. All these methods have their own set of weaknesses and currently there exists only very few solutions implemented by Google called the Recaptcha and the account limit by ebay. But with the advances in data mining technologies there is a chance in near future to evade even Recaptcha.

Email verification is one way to verify accounts but it can be overcome by using disposable emails that are offered by few websites/tools.

Similarly, phone verification can also be bypassed by using temporary numbers provided by some websites mentioned in the details section.

Facebook has come up with watermarking technology to identify the ownership of pictures. But it is known to fail when the picture is modified and save on different name.

Captchas are very ineffective as the tools solve them quickly and easily than humans. One method is to pose a math puzzle or ask general familiar questions. This can be overcome in sometime by generating database of these questions and answers and learning from them.

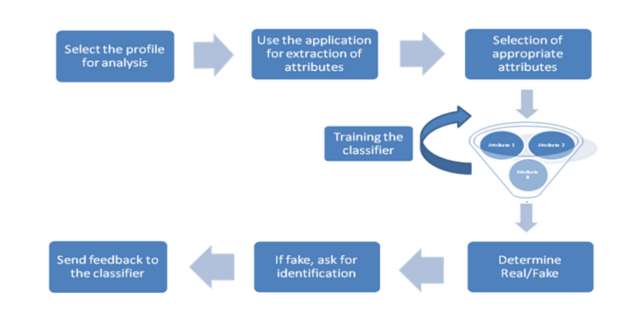
Limit on the number of accounts per IP per day seems to be the most effective up to now and its been successfully implemented by ebay. The mass account creator can always switch IPs by requesting new ones from proxy or VPNs but this can be tracked and the accounts can be suspended. This method do not halt fake account creation all together.

Recaptcha which analyses user behaviour and then poses various media challenges proves to be the best available method. This method also cannot stop the creation altogether if advanced learning techniques are used.

**How people fall for fake accounts.**

**Approach**

**Machine learning approach**



**Fig. 27 The Different Components of the Machine Learning Pipeline [39]**

To analyze different reasons that contribute to the fall for fake accounts, data mining which has to follow normal stages to achieve the goal can be of useful tool to implement prediction of whether some reasons will lead to a person fall for fake accounts.This requires sufficient data sample to be collected as well as authenticated data [39, 14].

About fake profiles. While the collection of data from real profiles was relatively easy we can collect them through existing friend circles, the challenge was in having sufficient number of identified fake profiles. The data mining approach takes use of several classification algorithm:

1) K Means:

The aim of the K-means algorithm is to divide M points in N dimensions into K clusters so that the within-cluster sum of squares is minimized. Only when M, N are small and K = 2 the solution can have minimal sum of squares against all partitions. We use the k means algorithm to judge the different reasons behind how people fall for fake accounts.We clustered some similar reasons into one group and analyze each group for their happening rates [39, 41].

2) Naive Bayesian Classifier:

It uses probabilistic methods to assign similar classes into one cluster and is used in fields of information retrieval. We used two algorithms for the classifier: Gaussian distribution, multinomial distribution. These two algorithms can draw distributed graphs that analyze how people fall for fake accounts [39, 2].

3) Support Vector Machine:

They are non probabilistic classifiers and used binary classification.SVM classify by using different features as different points in multidimensional space and then separates out the different classes using maximum margin hyperplanes.

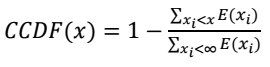
Distinguishing between real and fake users on online social network has been always a difficult and challenging task. Among them SVM(Support Vector Machine) proves to have better distinguishing rates and is more convincing [39, 21].

**Feature-based analysis**

**Implementation 1:**

This is an URL where users are allowed to upload their own image to personalize their account. And our analysis mainly focus on analyse the fake accounts URL and see what kinds of characteristics would make real accounts believe they are real accounts.

The complementary cumulative distribution function (CCDF) which is aimed at calculating the shannon entropy of x is computed as the following equation:



The fake group are seen to have similar URL among many profiles. This condition causes a small value of normalized standard deviation of entropies [6].

**Implementation2:**

Fake accounts are used for various reasons such as getting private information from users, sending spams to users forcefully or to enable people trust fake profiles for further abuse usage. People also create fake profiles for their acquaintances.They can do stalking,friendly pranks, cyberbullying,and concealing a real identity to get to their purposes. Many fake accounts are also created for getting benefits from online games [40].

Due to the variance of the reasons behind the creation of fake accounts, there can be multiple malicious behavior resulting from all these bad motivations. Attackers can avoid detection by classifiers by adapting their behaviors so that classifiers will not recognize them [40].

People fall for these accounts also for various reasons.Above is the feature for fake accounts and these features are diverse and unpredictable. In this way, the automated feature-based detection of fake accounts used by social networks can show very poor results [40].

**Coverage**

Rather than taking advantage of system vulnerabilities, fake accounts take advantage of

the way humans interact with computers or interpret messages, exploiting the difference between the system model and the user's’ mental model [14, 2, 6].

Fake (Sybil) online social networks accounts can be used for various purposes. For instance, they enable spammers to abuse an online social networks messaging system to post spam, or waste an OSN advertising customer’s resources by making him pay for online ad clicks or impressions from or to fake profiles. Fake accounts can also be used to acquire users’ private contact lists.Furthermore, fake accounts can be used to access personal user information and perform large-scale crawls over social graphs[2].

We have found that during all these purposes, most people fall for fake accounts because these fake accounts provide some attractive benefits. Under these benefits, people tend to believe the benefits without judging if these accounts are real or not. Another second most happening occasion is that some fake accounts may be opened by people’s friends. These friends, for some unknown purposes would use the information he/she already knows about the person to make the fake accounts appeal the person’s interest. In this way, one may feel quite familiar with the fake account and can easily drop their caution and fall for these account.

The last most used way to make people fall for fake accounts is to fake the accounts into someone really attractive. For example, handsome rich man or beautiful young woman. They fake their background and picture so that people will fall for them and tend to believe them because the identity they faked.

Fake accounts are created for profitable malicious activities, such as spamming, click-fraud, malware distribution, and identity fraud. Some fakes are created to increase the visibility of niche content, forum posts, and fan pages by manipulating votes or view counts.

People also create fake profiles for social reasons. These include friendly pranks, stalking, cyberbullying, and concealing a real identity to bypass real-life constraints. Many fake accounts are also created for social online games. For all these reasons, people are easily fall for the fake accounts and believe the accounts over there are created by real person.

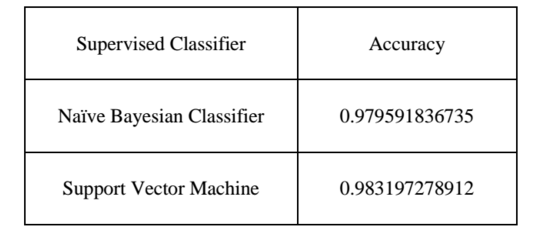
**Data & Analysis**

**Machine Learning**

In  *Audit and Analysis of Imposters in Online Social Networks*, the researchers studied data from 9097 total Facebook profiles, and used their supervised and unsupervised machine learning techniques to characterize whether or not a particular profile was fake or not. They included within that data set 3905 fake profiles and 5192 real profiles [39]. The goal of their experiment was to determine which of the techniques allowed for the highest accuracy of fake profile detection.

The primary difference between the techniques is in the unsupervised technique the data is not classified beforehand. Therefore the system has to classify the data first, which in turn will be used to train the classifier, which will then be used to test the data. While in the supervised pipeline, since the data is already classified, more of the data can used to test the data while simultaneously training the classifier [39]. It seems intuitive that the supervised technique will yield a higher accuracy since the data has already been classified, assuming that the classification is correct, it should yield to higher accuracy since it does not rely on classification based on the attributes of the data.

As expected the supervised technique was minutely more accurate than the unsupervised. For the unsupervised technique, the researchers used K-means clustering and the accuracy was approximately 96%. The supervised technique implemented Naive Bayesian and SVM. Table 1 below displays the accuracy results for the supervised technique, and it is evident that the SVM classifier provided the highest accuracy (98%) of all the aforementioned methods [39].



**Table 13 Accuracy Results for Classifiers [39]**

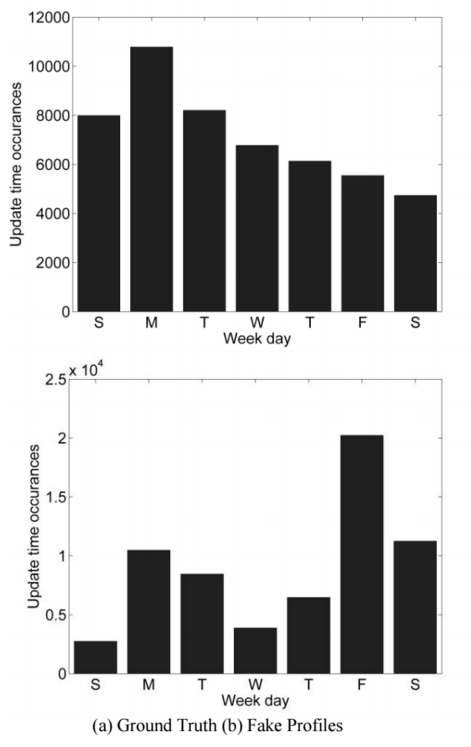
A potential issue with this approach is that it relies on publicly available information in determining the validity of an account. However, based on the user’s personal security settings, a lot of information may not be publically available. Therefore, the issue of false positives comes to mind, since their solution relies on the more information that is public the higher the probability that it is a real account.

**Feature-based/Activity-based Pattern Detection**

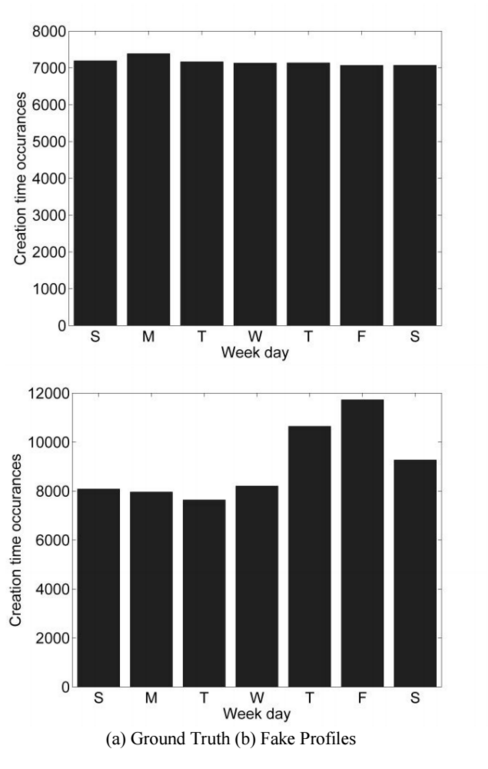
**Implementation 1:**

In *Fake Twitter accounts: Profile characteristics obtained using an activity-based pattern detection approach*, the authors chose a specific set of characteristics to monitor and compare other accounts to, so as to create a better machine learning pipeline. The real account data set is similar to the fake account data set in regards to size and timeline of tweets [40]. The research team used a Twitter crawler to create a database of user accounts, which numbered approximately 62 million [40].

Researchers found that there existed a difference in the distribution of the update times and creation times of accounts between the two different data sets. The results from Figure 3 show that fake accounts prefer to tweet later during the week, as opposed to real accounts which tweet more towards the beginning of the week. Figure 4 similarly shows that new fake accounts are created more towards the later part of the week, as opposed to real accounts. The authors suspect that a manual aspect of the creation and updating process might exist, thereby showing a preference for days towards the end of the week [40].



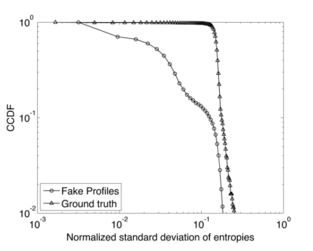
**Fig. 28 Update time occurrences vs. Day of the Week [40]**



**Fig. 29 Creation time occurrences vs Day of Week [40]**

Other interesting observations by the researchers involved the time differences between the two data sets. They noticed that fake accounts took a lot shorter time to create (20-40 seconds between accounts, though sometimes as fast as 3 to 5 seconds) and they were created sporadically during the day, real accounts on the other hand were uniformly created throughout the day and took longer to create. This suggests that fake accounts were created in batches [40].

The URL analysis component of this application traces how many different types of URLs being used for the profile picture of the accounts. Even if there are various different URLs being used, this application will be able to tell whether or not the picture is the same. The authors use a function called the complementary cumulative distribution function (CCDF) to measure the similarity of URLs. They discovered the CCDF showed that the URLS were largely similar for the profile pictures, and sometimes the same, as seen below in Figure 5 [40].



**Fig. 30 CCDF of the normalized standard deviation of entropies [40]**

A potential shortcoming the authors note is that their application was able to catch very few fake accounts. In hopes of keeping false positives down, they sacrifice some of their ability to maximize spam control, but for their application it is ideal [40].

**Implementation 2:**

The authors of *Aiding the Detection of Fake Accounts in Large Scale Social Online Services* argue that the feature-based approaches are not functionally viable due to the random nature of different user’s actions in OSNs. Due to the high variability in actions it is difficult to prevent a high false negative and positive rates, which leads to dissatisfied customers whose accounts get flagged as fake even though they are legitimate users. Balancing the need for targeting fake accounts and preventing the targeting of legitimate accounts is a real challenge, and if the false negative and positive rates are too high then the approach will not be feasible since it will require expensive human-level monitoring of the flagged accounts [6].

Another challenge for this type of approach comes from the adaptability of the attackers themselves. Once, they recognize the pattern of behavior that results in the flagging of an account, the attacker simply has to change his or her behavior to bypass this defense mechanism [6]. Therefore, the system of flagging certain behaviors has to be constantly updated based upon the type of behaviors that are routinely being used by fake profiles. However, as previously mentioned the unpredictability of user behavior makes this impractical.

The authors cite Facebook’s automated Immune System as an example of the impracticality of feature-based approaches. Facebook’s system only detected 20% of fake profiles. In addition, the fake profiles that were detected was because of the complaints of legitimate users [6].

**Strengths and Weaknesses**

Our approach relied on understanding the current scholarship on identifying why users fall for fake accounts, as well as how OSNs detect them. We researched three of the most popular approaches networks use to identify and stop fake accounts. The diversity of approaches, social networks, and applications allowed for us to see the challenges of quantifying and qualifying fake profile characteristics, behaviors, and most importantly their adaptability to OSN detection techniques.

The sheer number of OSNs and the variability of their protocols and standards resulted in a wide variety of information. It became difficult to create a consistent view of what types of indicators users and OSNs should be cautious of, however, for the purpose of this study it did demonstrate the reason why there are so many fake profiles on OSNs.

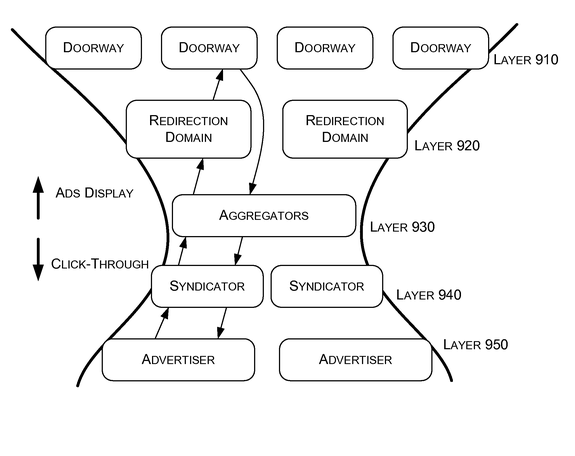
**How does information diffuse on fake accounts and the factors that boost the spread of information?**

**Approaches and Coverage-1:** Online social media is emerging as one of the important channels of information diffusion during real world events. An increasing amount of research is being conducted to examine the information flow in online networks. According to the past literature, Twitter and blogs are linked with faster diffusion of information. Most of these works focus on investigating the interplay between the different platforms, the kinds of events that influence where the information will diffuse first and the relationship between the type of an event versus the way information gets diffused. As the first step, multiple case studies are analyzed in this paper to understand how exactly information gets diffused in different scenarios. Also, we look at how fake accounts are created and play a key role in diffusing information during the planned and unplanned events. Alongside, we studied the various factors involved in terms of the spreading malicious or fake news. Especially, these fake news or rumours according to our research are prevalent during disasters or major catastrophic events. To address this, in this specific subsection, we first focus on the important approaches followed under different real-world scenarios especially how information contagion happens in different scenarios or events -- Boston Marathon, Hurricane Sandy, Wikipedia Hoaxes, Cyber criminal ecosystem on Twitter, political unrest, etc. Here are the main approaches and coverage of these different scenarios:

1. **Political Unrest:** During the political unrest, there are two types of people involved in the protests -- protesters who are involved in starting or leading the protest and authorities who are trying to control the protest. There are three main event types that can happen during the political unrest that can create distinctive information flow patterns:
   1. Events not planned by either protesters or authorities
   2. Events planned by authorities
   3. Events planned by the protesters
   4. Among the three events, during the surprise events that are not planned by either protesters or authorities and during the events that were planned by the protesters, Twitter activity appears before any online news media and blogs. Whereas, for the events planned by the authorities, the event was very well discussed before it actually happened. For events planned by authorities, the tweet volume quickly ramps up to a spike and then decays and goes flat as interest in the topic fades away. When the events are sudden, the volume of posts online are very high and there is a lag between the event and the beginning of the information sharing about the event. Planned events exhibit different information practices than sudden events. The activity on Twitter happens in real-time and the volume on this platform is very high and different compared to any other platform. In contrast, the other platforms like mainstream news source and blogs they focus on an event even days after an event ended.
2. **Boston Marathon:** Malicious content posted online can bring destruction and can create a huge damage especially when the events are related to disasters. Twitter being the fastest source of news, many malicious or fake accounts are created and fake news gets diffused on this platform very easily. The research on this event performed an analysis that includes data collection, fake content tagging, temporal and spatial analysis, user description analysis, to analyze the fake content and suspended profiles on Twitter. Results revealed that users with high social reputation and verified accounts were responsible for spreading the fake content. Closed community structure and star formation in the interaction network are found from the suspended account analysis.
3. **Hurricane Sandy:** This research focused on how Twitter played a role in spreading fake images during the disaster. A characterization analysis to understand the temporal, social reputation and influence patterns that help in spreading the information is performed. This research resulted in identifying that most of the fake information are retweets and surprisingly, network links such as follower relationships of Twitter played little or very less role in diffusion. This research also focused on classifying the fake images from real images and showed that this kind of automated analysis works well to perform classification and also decision tree based classifier is the best technique to perform classification in this scenario.
4. **Wikipedia Hoaxes:** Hoax articles do receive attention on Wikipedia even though there are moderators who try to control the hoaxes. In this regard, the research focusing on understanding Wikipedia hoaxes investigated the impact, characteristics and detection of hoax articles. Essentially, the main focus is on measuring the impact of hoax articles before they get disabled in terms of how long they survived, how many page views they obtained and how many references of these articles are made on the web. Results revealed that Wiki community is very efficient in identifying the hoax articles. Features like the article’s structure and content, embeddedness into the rest of Wikipedia and features of the author who created the account are very important.
5. **Cyber Criminal Ecosystem:** Empirical analysis on how the ecosystem of cyber criminals exist on Twitter is the main goal of this fourth research case study. Two types of analysis are conducted that includes the inner social relationships and outer social relationships that essentially focus on the users within the ecosystem and between a user in this ecosystem with his/her outside social network friends. By measuring this network density, reciprocity, etc., it was revealed that inner social relationships tend to follow a small-world network structure where multiple criminal accounts are socially connected. Outer social relationship analysis revealed three different categories of accounts -- social butterflies, social promoters and dummies, who closely maintain relationship with criminal accounts online. By utilizing the important features like social relationships and semantic coordinations, they developed a criminal account inference algorithm.

Here below, we focus on the different types of content that were used to detect whether given news is fake or real. To handle, this different real-world scenarios are considered along with the different types of content cues.

1. **Diffusion of messages in a social network:** We focus on message diffusion in social media using hashtag taking the example of rumor and non rumor tweets during the boston bombing of 2013. After the tragedy, twitter was filled with false messages propagating confusion among the people and to mitigate this more messages carrying correct information was needed. In order to stop this , all messages were analysed and categorized based on hashtags into categories and the two main sources and their followers were identified through message origin and reaction time [24].
2. **Account hijacking on twitter:** One of the biggest consequences of the above mentioned attacks is account hijacking, the next paper deals with this kind of hijacking in twitter.The research focuses on the consequences of Account hijacking on users and web services primarily Twitter. It was found that 21 percent of the victims never returned to use the service despite complete recovery by Twitter. Malware and spams are the biggest source of hijacking in Twitter. Biggest challenge is distinguishing fraudulent and victimized accounts. To identify accounts we cluster the data, classify the cluster, crawl across connections and identify spams, this is achieved using Hadoop, Pig or Spark. It was found that if Twitter identifies fake accounts within the first 24 hours, it could save 70 percent of the victims. Classification of users is achieved through metrics such as followers, following, tweets, maturity and language [18].
3. **Spam URL detection:** This concept is built above the concepts of the previous paper of predicting information in Twitter URLs. It is found that 8% of 25 Million tweeted URLs are found to be malwares, spammers and phishers, in order to prevent damage, early detection is vital as even after blacklisting most of the users are already affected. A faster mechanism is proposed in this paper. Spam accounts are classified as compromised and career spamming accounts. Only 50 percent of spams can be detected using repeated keywords. So the proper mechanism is to cluster URLs into campaign data sets. Using Cluster data , we identify the Blacklisted accounts that spread malware or spams taking into consideration retweet rate, nested URL shortening( which poses a major problem in identification of Spam URLs). Blacklists are contingent on domain of the user and hence multiple domains renders this concept ineffective. This system allows us to distinguish bots from compromised accounts and also leads us to the end spamming page. Due to the enormity of the number of URLs detection of Blacklist is delayed, which causes the spammers to gain millions of more spammers, this is a problem which has not been solved here [27] .
4. **Impact of the # hash Tag:** Hashtags have evolved to become the bookmark of content on social networks especially twitter. Data on twitter can be easily classified using the hashtag and their content categorized using simple mining techniques. This system proposes a model by which we can predict the use of hashtags by a user even before that hashtag have ever come into use. Experiments are conducted on real time data acquired from twitter. One major constraint is the variation of relationships between the user such as friend, colleague and hence when they are mentioned and retweeted, the relationship has to be identified in order to better understand the content purpose. The hashtag is considered as the pointer to content and a measure of the interest of the person. Preference of hashtag varies between individual users and hence is a unique measure of user style, which is determined through previous content of the user. The community the user joins also majorly influences his contents and hashtags. Regression analysis is carried out on the data set acquired, ignoring a few negative cases that were observed the SVM system predicted the appropriate hashtags, this system could further be enhanced as a recommender service. This is left for future work to be built over the proposed system [36].
5. **Trust what we retweet ?** The behavior of the twitter users during an emergency situation taking the example of Chile Earthquake of 2010 is the primary focus here. The spread of news, rumors across twitter and the immediate propagation of these tweets, their reliability is studied in depth. Analysis of the social network activity during the occurrence of the incident was studied based on categorization of tweets in accordance with the keywords. For example during an earthquake the keyword earthquake is used to retrieve the relevant tweets and the tweet data were mined for terms that were most commonly associated with the given term and they were tabulated and analyzed. Based on this analysis it was identified that false rumors were not accepted immediately than expected, while truths had more response because the rumors were questioned more by the community and hence they act as a collaborative filter and aggregate analysis could hence detect rumors. In addition twitter was also used for helplines, emergency evacuation strategies , which dramatically improved response time plus could cover a wider area than any other form of communication [30].
6. **Spam detection in twitter:** Twitter offers Spam detection features to user, where in the user can report a particular tweet or user as spam but this in turn is misused by spammer making them ineffective, also the manual reporting depends on the experience and expertise of the user. Hence this paper [31], proposes the building of a social graph based on followers, friend relationships. Classification strategies are used to identify suspicious activity from others and using Web crawlers, the Friends and followers information is retrieved for 49 million users which is used to draw the graph. K nearest neighbor algorithm is used atop Bayesian classification to remove noise and predict spam accounts and tweets. However it was found that the output reported very less actual spam when using K neighbor alone, while a 89 percent result when Bayesian was used. Retweets are identified by the system as links which lead to the spam origin and other spammer accounts. The F measure for Bayesian classification which is a measure of Precision and recall rates was found to be F = 2PR/(P+R).
7. **Signatures and characteristics:** AutoRE, an automated URL signature generating framework is introduced here to identify spam emails [29] . Bots are said to be behind most spam emails. AutoRE extracts URL Strings, email sending time and source server IP from a given email discarding forwards ( reducing the data set) and partitions the emails into groups based on domains and identifies the origin of the spam campaign. Due to the grouping the search cost is significantly low. Similarity of sending time, email properties and behavior are used by AutoRE for the detection process. This opens door for future spam diffusion detection and improves the false positive rate. This system assigns a unique signature using a regular expression, of two types namely complete URL and regex based signatures. It was found that botnet spam emails followed a specific pattern based on Gaussian model, hence they could be grouped into superclusters and hence could be identified early, however there is continuous evolution of botnet technology. Hence in most cases AutoRE acts as the post mortem for spam detection but has the future to work real time based on live feed, although spammers may also evade AutoRE by crafting URL based on selection process identified. Despite the above challenges, the quick signature generation feature could be extended for deployment in the near future.
8. **Understanding Spammers:** Characteristics of spammers based on IP pertaining to a particular region, time and host types are studied in this paper. Certain IP address ranges were found to have the most spam, this was identified through a study of 17 month trace of 10 million spam messages. According to the experiment it was found that most IP s used to spam were from Asia and categorized that spams could be were Direct Spamming, Open relays and Proxies, Botnets, Spectrum agility. It was observed that the Bots were never visiting the same domain twice and hence are getting more advanced with time. Network level behavior of spammers were studied through the acquired data and comparing them with legitimate email were easily able to identify spam messages. The ineffectiveness of black list due to its slow adoption was reassured in the findings of this survey as well and they proposed a scattering technique known as spectrum BGP agility that identified IP blocks but further encountered certain drawbacks in identifying individual accounts [32] .
9. **Spam Double Funnel:** Advertisers make use of spammers to advertise their products in search results or spam links such as clickjacking. This paper proposes a five layer model as shown in **Figure 31,** to show how this works from end to end. This involves two important categories of people, targeting spammers and advertisers. Strider Search Ranger system proposed here analyses the redirected traffic to known spam domains and detects spam. Keywords such as Drugs and ringtone have been identified to be the most common with spams. The five layer structure is , Layer 1 – the doorway domain where the user is attracted to selecting the redirection link, the layer 2- intermediate redirection domain, the 3rd – the aggregators who funnel the redirection to the 4 th layer the syndicator who actually do contain the advertisements influenced by layer 5 – the advertisers. Spamming in these low quality pages is much cheaper and has a better reach for the advertisers, hence they have been found to be on a phenomenal rise despite prevention strategies being adopted [28] .



**Figure 31**: Structure of a spam double funnel [28]

1. **Credibility analysis in social networks:** The credibility of a particular information perceived by the users, along with the factors influencing credibility is the focus in this paper. Two experiments manipulating the tweets in order to observe the credibility change from user’s point of view, were conducted and compared to what was currently given by search engines. Also what authors could implement in order to improve trust and credibility in their network was identified. Since search engines like Google and Bing have started to include twitter and facebook information in their search result, it is critical that these social networks be devoid of spam and false information. Based on a survey conducted where users were asked to determine the credibility of certain tweets and authors, it was identified the results were poor on the user’s side. That is the users were not able to identify false information and were influenced by an author based on author’s bio information, followers, retweets and trusted authors without even actually knowing them. Authors who did not have the default twitter icon were trusted that the other, such decisions were also made pertaining to the content, language and activeness of the author [33] .

**Strengths and Weaknesses -- 1:**

One major impact of diffusion in social networks is the viral spread of rumour information, botnets, etc. Despite the development of frameworks and systems to identify spam information, botnets etc, credibility of information can never be authenticated by a system without manual intervention. The truth behind any news can never be authenticated or understood by a system and so the current systems can incorporate human-in-the-loop kinds of interventions. Another drawback is the absence of a generic algorithm that is able to achieve early detection of spams and botnets. There are many systems proposed in the above papers but none devoid of shortcomings, hence development of a universal algorithm is essential.

**Approach and Coverage-2:**

In this section we focus on the different types of approaches that the current state-of-the-art literature proposed to identify the rumors and spread of fake news. We widely categorize these methods into four different types of approaches.

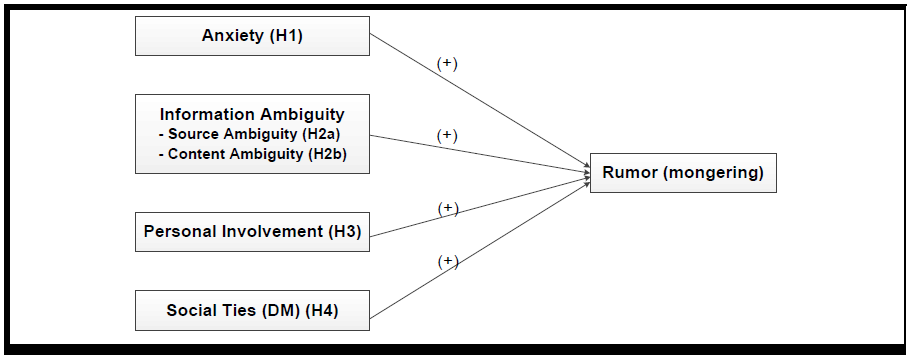
**-- Attack simulation-based approaches**: Literature shows that information diffusion depends on the scenario and network setting and in this paper we explored multiple case studies with respect to this. One of the popular works proved that deceptive attacks on social media are very virulently dependent on likelihood of a contagion effect. By simulating a phishing attack called as Farcing on Facebook, it was shown that there are two stages of attack. In the first stage the attackers or fake users use a phony profile to friend victims, and in the second stage, these fake users solicit personal information directly from victims. SImilar case studies proved that users who fall prey for fake users rely primarily on the number of friends or the picture of the requester as a heuristic cue and made snap judgments. Also, the diffusion of information from this fake user draws real users as friends because if one real user became friends with a fake user the other real users gravitate towards becoming friends with the fake user. Also maintaining more number of friends also attract more real users as friends. Such profiles caused an upward information cascade, where each victim attracted many more victims through a social contagion effect. More details are here below:

1. Contagion effect is the biggest perils of deceptive attacks on social networks. Close to one in five fall victim to farcing attacks on social networks and the rise in the number of victims is phenomenal. Judgement of individuals is dominated by number of connections, similarities, popularity and posts. Furthermore, when one victim falls for a farcing attack, it leads to a contagion effect threatening everyone in his network. College students were exposed to real time farcing attacks through fake accounts and the results were studied .Future research would include a broader study of individuals of varied age groups and not restricted to college students [19] .

**-- Scenario-based approaches**: Multiple real-world scenarios have been investigated in this study especially Twitter as the main platform. Initially, Twitter as a platform offers spam detection to its users but even in this case, there still are issues with fake users and fake news spreading virally. Psychological studies proved that fakeness takes a twirl especially during the unfortunate scenarios like natural disasters because of the three main factors. These factors revolve around issues like news or information with no clear source provided, personal involvement and anxiety. It has been revealed that the theme of the story along with psychological incentive for people to pass along the story are very important. Also, these case studies have revealed that rumor spread is a function of importance and informational ambiguity. More details are here below:

**-- Learning-based approaches:** Multiple regression, decision-tree based, support-vector machine related algorithms have been proposed to analyze the different factors and how they are statistically related to the main dependent variable -- information spread. Most research proposing these learning based approaches revealed that social relationships, semantic coordinations, temporality of how long the information has survived, measuring the credibility of this information by detecting the citations or references made to this fake news, etc., are thoroughly studied to measure how far and how wider the fake news spread online. More details are here below:

1. The focus here is on the diffusion of information in social networks by identifying the influencers who spread critical information rapidly across the network. Examples of this were identified in Facebook and Twitter during disasters. Detection of popular topics is critical, hence an automated system that identifies the current trending topics was developed. In order to improve accuracy and avoid mistakes, bursts was the primary focus, successive bursts were identified and the popularity was a measure to identify the temporal popularity of a topic across the social network. Graph based predictive model for this data predicts how the diffusion process takes place across the networks, two approaches graph and non-graph were discussed and the Bayesian classification was deployed for this purpose. An open source tool names SONDY could mine date in online social networks and implemented the techniques in this paper [34]. A social media campaign was identified to be optimized by targeting influential individuals who could trigger large scale cascades of adoptions. Using probabilistic clustering over this feature space , the nodes within the cluster were ranked to identify the most influential individuals. Topic definition, data complexity and social dimension are constraints that were considered during this analysis.
2. This paper [25], defines a framework that identifies the diffusion of memes in microblogs.It aims to identify political memes and astroturf political accounts within small diffusion networks in order to prevent them from gaining negative attention of the public.Early detection of these accounts is achieved by the proposed data mining framework.Using more views such as age, history of the user more analysis such as sentiment analysis, likeability, categorisation can be achieved.
3. The standard models used in the study (for example, studies like [46,47] of virality of information that is extended from the field of health sciences is called the *contact process* or the *susceptible-infected-susceptible (SIS) model*. In this model, every vertex is either infected or healthy (but susceptible). An Infected vertex becomes healthy with rate 1 independent of the status of its neighbors. A healthy vertex becomes infected at a rate equal to the propagation ratio of the disease, **𝜆** , times the number of its infected neighbors. The other modeling approaches include identifying the different factors ranging from the user-level to network or structure-level and utilizing them to build a generative models or other learning models [63] based on support vector machines for effective classification of fake posts versus real posts. Some of the features can be language related like type-token ratios, lexical entropy, error features, n-grams, syntactic features, semantic features, user related like biography, frequency of posts, social circle, friends, followers, network related features like subgraphs, cascades, etc.
4. Apart from the automated computational approaches and methodologies, studies [42] exist which consider the rumor theory (a social psychological perspective [64]) to explore the collective social reporting as an information processing mechanism to address crisis problems and gather community intelligence. Similarly theory has been utilized to explore the degeneration of social reporting into collective rumor mills. Rumoring is defined as a collective and collaborative transaction in which community members offer, evaluate, and interpret information to reach a common understanding of uncertain situations, to alleviate social tension, and to solve collective crisis problems. It was revealed that rumor spread is a function of *importance* and informational *ambiguity*. This explains that the theme of the story along with the psychological incentive for people to pass along the story are important. Through the analysis of Twitter data, it has been revealed that these are the three main rumor causing factors (among the four factors shown in **Figure 32**):
   1. information with no clear source provided,
   2. personal involvement and
   3. anxiety



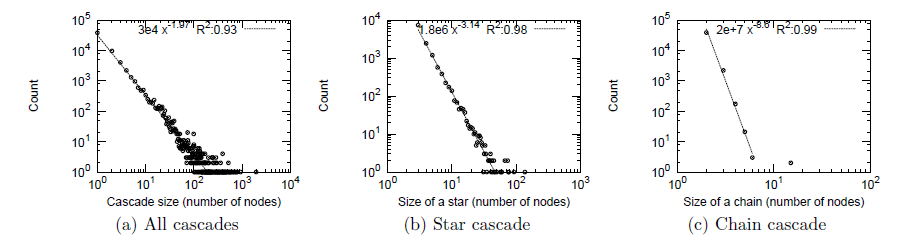
**Figure 32:** Different factors in which rumors spread and are positively correlated [42]

**-- Diffusion models:** Influence or the information flows or propagates like flu virus which is the SIS model

in epidemiology. Research in this sub-group has shown that the size of the cascades, size of blogs and other topological patterns follow a power law with basic components of cascades as stars and chains. As Influence propagates like SIS model in epidemiology, by understanding the deviations of patterns from these observed behaviors can help identify how the rumors or false information spreads online. More details are here below:

1. Influence, spread and rate of propagation of pictures across the social network Flickr is studied in depth here. Experiments were conducted to identify the propagation of popular and unpopular pictures across Flickr and also the impact of friend information exchange influences favorites. Favorite markings , Links, Users, Photos and time period were studied from a given real time data set of Flickr acquired through crawling, however favorite marking did not include deleted ones. Ranking of pictures were not only based on raw values of metrics such as number of views, fans and comments but a correlation between these values. The correlation however was inconsistent and never generated a pattern. Hence other factors influencing popularity of pictures were identified such as Local versus global popularity , distance from fans to picture uploaders and rate of growth of popularity of pictures and users. Social cascade hops play an important role in this concept. Due to bookmarking of friends photos even popular photos took a long time to diffuse through the network. A major application of the results of this study can be used to understand viral marketing across digital content in social networks [35].
2. Information virality is an important topic that is drawing increasing attention from researchers in different fields to investigate and understand factors that influence the process of virality. Information virality is the process that gives any information report the maximum exposure relative to the potential audience over a short period of time. Theories of information diffusion mainly focuses on two diametrically opposite approaches to viral information diffusion:
   1. virality as a process governed by reliance on centrality of nodes in a given graph (as networks are structured as graphs)
   2. others argue about this process as a much more dynamic bottom-up process where the central or key nodes are important but not crucial

Multiple directions of research in multiple fields has been explored in terms of how virality happens. Shawn et. al, designed a new methodology in information science and communication borrowed from economics that allows to perform a time series analysis through multiple regression. This paper that addresses virality with respect to politics [39] has specified that information diffuses mainly with very popular actors not necessarily political actors. Researchers started exploring virality from the perspective of blogs and essentially how they cite and influence each other [40]. Research showcased that the popularity of a post online drops off with a power law distribution (shown in **Figure 33**) instead of dropping exponentially with time. Size of the cascades, size of blogs and other topological patterns follow a power law with basic components of cascades as stars and chains. Influence propagates like a flu-like virus that is, the SIS model in epidemiology. By understanding the deviations of patterns from these observed behaviors can help identify how the rumors or false information spreads online.

**Figure 33** The size distribution over all cascades (a), only stars (b), only chains (c). They all follow heavy tailed distributions with increasingly steeper slopes [40]

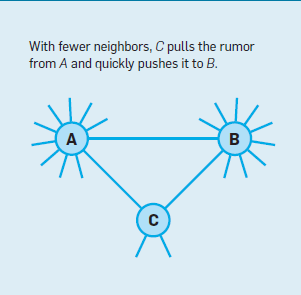
1. On the other hand, past research [41] validated that there are differences in the mechanics of information diffusion across topics in terms of idioms, hashtags, etc. Especially, there is a significant variation in the way different widely-used hashtags on multiple topics spread. When there are politically controversial topics, the hashtags are particularly persistent with repeated exposures to have unusually large marginal effects on adoption validating the complex contagion principle from sociology. Whereas some natural analogues of idioms and neologisms are particularly non-persistent with the effect of multiple exposures decaying rapidly. This may explain that in terms of rumors or fake news trending on Twitter, information may decay rapidly but the extent to which the information spreads may depend on the topic. Also, studying the subgraph structure of initial adopters and finding the structural differences of any post or news trending online can be used as one of the features in classifying rumors. Wide range of simulation-based and generative models have been developed to analyze how the dynamics of adoption interact with the network structure of the early adopters.

**Strengths and Weaknesses -- 2:**

Most of these approaches focus on studying the rate of diffusion on Twitter using a very limited dataset. which significantly restricts our knowledge of diffusion on a bigger scale as we do not take into account characteristics of users of different ages, ethnicity, diversity. Also, the diffusion modeling approaches leveraged by not considering the external influences when some information is getting diffused. Future work should consider into account the diversity and the sub network that replicates the original network that takes into account all different types of users on the network. Faster performance is another key factor to be considered as by the time the fake accounts are detected, most of the victims are compromised. In the future, systems should be designed to achieve both faster performance and cover a wider set of people.

**Important Insight: Why non-trustworthy news spread so quickly in Social Networks**

To analyze how fake news spreads in social network, this study [44] simulated a simple information-spreading process in various network topologies. This analysis demonstrated that news spreads much more quickly in existing social-network topologies than in other network topologies. The main interesting scenario is that even though social networks grow organically, their graph structure is not designed for any particular use. But still this structure allows for the quick spread of news.

  
 **Figure 34** How a rumor progresses from a large-degree   
 node A to another node B; due to their having large number of neighbors [45]

The main source of the quick spread of information is the truthful interaction between the few nodes with many neighbors and the large number of nodes with few neighbors. Small-degree nodes learn a rumor once one of their neighbors knows it, then quickly forward it to their neighbors. This propagation scheme facilitates sending rumors from one large degree node to another. As shown in **Figure 34**, nodes like C with fewer neighbors quickly pulls the fake information from A and pushes it to B where A and B need more time to push or pull the rumor. The push-pull strategy to deliver the message to all nodes with certain complexity is explained clearly in this [45] study.

**CONCLUSION**

**Comparison of Strengths and Weaknesses**

**How to detect or identify fake accounts and predict friends vs. foes.**

The Facebook Immune System posits that it will be sufficient to protect against fake accounts. The paper does not support this claim with any objective evidence. This makes the claim immediately suspicious. The Facebook Immune System may be theoretically secure. However, theoretical security does not guarantee actual security. The paper *The Socialbot Network* shows that the Facebook Immune System is insufficient to protect the system against fake accounts.

In *The Socialbot Network* the authors created an automated system for generating fake accounts. The paper outlines how to automate a fake account creation process and how to bypass fake account creation obstacles. The system had minimal cost to run. To reiterate, out of the 102 fake accounts running for 8 weeks, 20 were blocked by the Facebook Immune System. These accounts were blocked because of user reports. These accounts were not flagged as fake accounts, but rather were flagged as spam. Of the overall attack 70% of fake accounts succeeded in connecting to 0-50 real accounts. In addition 23% of fake accounts succeeded in connecting to 60-80 real accounts. The system was able to collect an average of 175 chunks of new personal identification information each day. [3]

These statistics show that the Facebook Immune System failed to correctly classify classify the accounts in the experiment as fake accounts. Only 20 of the fake accounts were blocked by the Facebook Immune System over the course of 8 weeks. Furthermore, because these accounts were reported by users, this means that the Facebook Immune System’s dynamic detection feature did not succeed at detecting any part of the fake account network. This shows that the Facebook Immune System does not detect the fake accounts, regardless of what feature of the fake account is detected.

In addition, the Facebook Immune System did not classify these accounts correctly. This provides even stronger evidence that the Facebook Immune System does not detect fake accounts. This is in contrast to the lack of experimental evidence provided in the *Facebook Immune System*.

Furthermore, the fake accounts were able to obtain potentially useful information. The fake accounts succeeded in connecting to real accounts. Not only did they connect to other accounts, but they also succeeded in retrieving personal identification information. This means that a malicious attacker could gain personal identification information by using an inexpensive fake account creation method. This would be a significant failure of the intent of the Facebook Immune System.

With this strong evidence against the efficacy of the Facebook Immune System, it would be reasonable to conclude that the current system used by Facebook is ineffective. One cannot recommend that the current system be used for fake account detection. Therefore, a better system is required.

Detection of clusters of fake accounts is best method compared to fake account detection in twitter based on minimum weighted feature set because in case detection of clusters of fake accounts the detection happens in the cluster level where the cluster is the cluster of accounts. The main advantage of following this approach, the detection of fake accounts will happen during the time of the registration of accounts this means the a fake account is not allowed to enter network and perform some actions such as connecting with friends, sharing, commenting on posts. But in the other method, i.e. Fake account detection in twitter based on minimum weighted feature set, will work only when a person allowed inside the network and perform some actions. Moreover, this method of detection is not generalized to all OSN and it is restricted to only twitter. So, I strongly conclude that detection of clusters of fake accounts is better in terms of the time taken, complexity and it is generalized method of detection across all Online Social Network.

Taking a different approach from the traditional machine learning based approaches to detect fake accounts in OSNs, SybilRank uses power iterations to perform short random walks in the social graph, starting with a few selected trust seeds (legitimate accounts). The whole idea of the system is based on the assumption that Sybil accounts are very sparsely linked to non-Sybil accounts as social relationships need to be bidirectional, and this would produce a sparse cut in the social network graph. This system basically aims to concentrate human efforts to only those accounts which are highly likely to be fake accounts rather than going through all of the user accounts manually, which is very time consuming and also cost ineffective.

The paper describes the experimental results of implementing the system on Tuenti, one of Spain’s largest social networks which used to employ manual effort. The paper also highlights how SybilRank can perform more effectively, with lower rates of false positives and false negatives at the same time be cost effective and have higher fake account capture. It describes that the social network is a densely connected graph of nodes which represent user accounts. There are two kinds of accounts- legitimate and malicious. The legitimate accounts usually are densely connected to multiple other accounts as expected in a normal social network scenario. Exceptions to this assumption can be newly created accounts or accounts that are not very active. When it comes to the Sybil accounts, they tend to be isolated in a way that they do not have many connections to non-Sybil accounts. This is because maintaining a connection in an OSN is a bidirectional relation.

SybilRank starts with a few selected legitimate accounts (known as trust seeds) and performs short random walks to all the accounts linked to the trust seeds. The system allows any number of trusted accounts to be considered as trust seeds. In order to make the system robust, we may start with a bigger number of trust seeds. A random walk simply performs a graph traversal starting from non-Sybil trust seeds and stops when it encounters a break in the graph. A break in the walk or a sparse cut in the social graph is an indication of entering a Sybil region of the graph. Each account is then assigned a score which depends on it’s interconnections from the trust seeds. Finally SybilRank produces a ranked listing of all the accounts, with the highly likely fake accounts at the lower end and the less likely ones at the top. The human analysts can then start analyzing accounts from the lower end of this list and about ~90% of the times, these accounts turn out to be fake.

The ranked list approach used by SybilRank helps reduce human effort and is effective even in large OSNs. However, it might wrongly identify inactive accounts or new accounts as Sybils too. Compared to several machine learning based approaches, SybilRank does perform better.

When it comes to COMPA system for detecting compromised accounts, an effort has been made to curb the accounts that are compromised but are still densely connected to other accounts in the social graph as a legitimate account has been taken over in this case. Such accounts would pass undetected by a system such as SybilRank or by many other ML-based approaches.

By creating a behavioral profile for all the account users, based on a set of certain selected user features, any deviation from the normal, expected behavior of the user can be detected which would strongly indicate that the account has probably been compromised. There are quite few systems that work towards detection of compromised accounts, and the COMPA system does serve the purpose quite well.

For the approaches mentioned above, the first approach, graph-based approach, which is almost completely based on the fake accounts feature and people psychology feature when they face fake accounts to do the analysis. Thus some of the data are varying from a big range and can result in many uncertainty. Also real users response to fake accounts can sometimes be ambiguous and hard to decide their real attitude towards these fake accounts. What’s worse is that even if people fall for fake accounts, there can be multiple reasons on how they fall for them thus making it even harder to only do graph-based analysis.

However, this approach does have some advantages that it can take all the possible reasons into account according to each feature. This can make the conditions more applicable and give some hints on how people fall for fake accounts. Sometimes some reasons need a combination to mix analyze. When this occurs this can cause some possibilities to mix some reasons. With graph-based approach, mix reasons can have more validity and make more sense in combining.

**Studying the network and behavior of fake accounts**

An advantage among majority of the discussed methods is that they are good at detecting groups or ‘clusters’ of fake accounts, which is exactly the kind of scalability that is expected when dealing with the magnitude to today’s social network users. But very few methods, like ‘clustering’ [2], are concentrating on the time sensitiveness of the fake account problem, which is to catch these fake profiles early in the system.

Almost all of the methods need manual intervention at one stage to ascertain whether an account is fake or not (clustering [2] has both automatic and manual actions). While an automated system is desirable, the network administrators must be so conscious not to accidentally block or remove any account, and hence the manual part becomes a necessary inconvenience.

Most approaches or methods we have seen for detection of fake account networks suffers from one major drawback, that it is mostly limited to work within 1 or 2 online social networks (identity clone attacks [11], profile cloning [10]). There has not been a general approach yet that can cater to the wide range of different OSNs that we have today. Understandably, the diversity of the purpose of different OSNs make it a difficult task to generalize the information coming out of them and hence the challenge in having a common approach or generalized form of data to analyse.

The above methods suffer from not being able to be tested on a large scale in the real world, as creating a huge number of fake accounts in real world for testing might damage and compromise the system.

Coming to the Privacy leakage of normal users to the fake account users[8], being a case study and not a method, there doesn't exist any comparisons with other methods. However we want to mention the following conclusion and how the case study can be improved. We studied assessment of the seriousness of privacy leakage problem by creating eight fake account users and conducting series of experiments. Then we studied how quantitatively the of degree of online self-disclosure (DOSD) is analyzed and how the photographs of legit users are leaked. Thereby, we studied the report of possible risks of privacy leakage and discussed possible ways of mitigating the problem. However, there is only the discussion of only some users in online social networks, the self disclosure or privacy leakage of large scale users will be a valuable problem.

In Multiple Identity/Fake Account detection using nonverbal behavior we discussed[7], the non-verbal behavior analysis is a good alternative path, which can be able to use to detect fake users. Whereas in Anomalous user behavior in Online Social Networks using PCA analysis, it gives more systematic with general framework for studying the user behavior. The real data from different social networks are taken to demonstrate legit user behavior with a set of features/dimensions given to PCA. There is also evaluation of the results based on ground-truth data with achieving good detection rate of fake account users.

Coming to the comparison of weaknesses, In Multiple Identity/Fake Account detection using nonverbal behavior we discussed[9], there should be coordinated effort to test this method in different platforms. The efficiency and effectiveness of this method depends on many context specific factors. The time window for observing user behavior has significant impact on the effectiveness of this method. The size of the window may also have an impact on effectiveness. There is also need of identifying what exactly are the non-verbal behavior variables in online social networks. Whereas in PCA analysis,to achieve more detection rate of fake users when combines with some supervised learning strategies. PCA analysis may not be ideal for small crowdsourcing online systems, where we can chose supervised learning.

**What attributes lead to the success of fake accounts?**

In fake account creation context, we can we see that many websites do not implement stringent security measures and most of the measures known can be bypassed by existing tools. Let us discuss some countermeasure methods to combat these fake account creation practices.

Most of the available tools for fake accounts creation are very effective. Additionally, these tools are widely available making it even easy for the attackers. As a countermeasure method, we can identify these popular tool websites and block them. Even if the number of accounts per IP are restricted, this can be easily bypassed by fetching fresh IPs from VPNs or new IPs can be rented from botnets. One way to counteract this is to limit number of accounts per IP per day. This would force the attacker to request new IPs from the free proxy and VPN providers allowing ISPs to identity the misuse of these IPs and report the fake account activity.

On the other hand, it is relatively easy to tell if an email address used for account verification is fake. The fake accounts have long unusual ids and domains that originate from temporary email providers. We can ban any accounts that are verified using these domains.

Some websites look at few characteristics that could determine if the account is fake or real which include finding if there is sudden increase in number of friends, followers or likes, minimal account activity, profile containing spam content,real users markin persons as unknown, profile has other fake users as friends. etc.

In the social engineering context, studies proved that having mutual friends improves the likelihood of accepting a connection request from an adversary. Most users often logon to OSNs and approximately half of the accepted connection requests are within a day of the requests being sent. Users with more number of friends are likely to accept requests from strangers (adversaries).

All these attributes need to be studied and defenses against these attacks must be developed to protect users’ privacy and the integrity of data available on OSNs. These studies help OSNs to design or improve techniques to detect fake accounts and make it difficult for attackers to create fake accounts.

**How do people fall for fake accounts.**

**Machine learning approach**

The data mining approach uses the dataset as input to analyze how people fall for fake accounts. Different algorithms of data classification can generate different results. Also this approach depends much on the validity of the input dataset and the attributes and labels it has. Some dataset with good sample data and reliable attributes can have better results under same classification approach.

We used K means, naive bayesian classifier,support vector machine these three methods as the classifier. These three methods can analyzing different aspects of how people fall for fake accounts.K means, which can cluster similar reasons and cluster similar sample data as one group. This method is good at using current sample data to generate good cluster. In this way, we can better group the reasons in how people fall for fake accounts and avoid unuseful attributes to raise the accuracy.

After using K means to cluster the data input, we then used naive bayesian classifier to draw the distribution graph. Gaussian distribution and multinomial distribution were used for analysing the distribution. These two distributions are currently the most accurate ones in analysing large dataset. In this way, we maybe able to get the reasons where people fall for fake accounts the most and where the least.

Support Vector Machine is also a way to separates out the different classes using maximum margin hyperplanes. We mainly use this method as a counterpart of K means to prove if K means give out right results. This method may be better than K means in that it is non probabilistic classifiers and used binary classification.This can largely reduce bias of different dataset and can generate more reliable result.

Logistic Regression makes use of parameters which uses maximum likelihood criterion.The key point in this method is to find the tradeoff between performance and overhead. I f the tradeoff is maintained, this method can have good performance.Random Forest combines several weak classifiers into one strong one and is stronger than those weak classifiers.

The machine learning pipeline, which is three step pipeline includes all the analysis, practise and prediction steps. This whole flow is complete and valid. This pipeline also uses above data mining methods to make the method more valid.

**Feature-based**

It is kind of similar to first one in that it takes into account what kind of features will make people fall for fake accounts. However this one may be more valid than first approach in that it uses real data analysis and set URL as a measurement of features. This URL measure can be more visualize than simply analyze all possible features.We also computed the normalized standard deviation of the entropies in each group to see how the reasons are gathered together to make people fall for fake accounts. For fake accounts, even when their URLs were very different, the images were seen to be the same. A large fraction of profiles in the fake group were actually seen to have similar or the same URL, resulting is very small values of normalized standard deviation of entropies.

One limitation of the approach is that it only identifies a relatively small percentage of fake accounts. But the low number of false positives that are likely in the obtained fake profiles make it an ideal seed database for use with social graph techniques for efficient spam detection.

**How does information diffuse on fake accounts compared to the real accounts?**

Different approaches that we have investigated in this paper can be categorized into three distinct high-level categories:

**Attack simulation-based approaches**: Literature shows that information diffusion depends on the scenario and network setting and in this paper we explored multiple case studies with respect to this. One of the popular works proved that deceptive attacks on social media are very virulently dependent on likelihood of a contagion effect. By simulating a phishing attack called as Farcing on Facebook, it was shown that there are two stages of attack. In the first stage the attackers or fake users use a phony profile to friend victims, and in the second stage, these fake users solicit personal information directly from victims. SImilar case studies proved that users who fall prey for fake users rely primarily on the number of friends or the picture of the requester as a heuristic cue and made snap judgments. Also, the diffusion of information from this fake user draws real users as friends because if one real user became friends with a fake user the other real users gravitate towards becoming friends with the fake user. Also maintaining more number of friends also attract more real users as friends. Such profiles caused an upward information cascade, where each victim attracted many more victims through a social contagion effect.

**Scenario-based approaches**: Multiple real-world scenarios have been investigated in this study especially Twitter as the main platform. Initially, Twitter as a platform offers spam detection to its users but even in this case, there still are issues with fake users and fake news spreading virally. Psychological studies proved that fakeness takes a twirl especially during the unfortunate scenarios like natural disasters because of the three main factors. These factors revolve around issues like news or information with no clear source provided, personal involvement and anxiety. It has been revealed that the theme of the story along with psychological incentive for people to pass along the story are very important. Also, these case studies have revealed that rumor spread is a function of importance and informational ambiguity.

**Learning-based approaches:** Multiple regression, decision-tree based, support-vector machine related algorithms have been proposed to analyze the different factors and how they are statistically related to the main dependent variable -- information spread. Most research proposing these learning based approaches revealed that social relationships, semantic coordinations, temporality of how long the information has survived, measuring the credibility of this information by detecting the citations or references made to this fake news, etc., are thoroughly studied to measure how far and how wider the fake news spread online.

**Diffusion models:** Influence or the information flows or propagates like flu virus which is the SIS model in epidemiology. Research in this sub-group has shown that the size of the cascades, size of blogs and other topological patterns follow a power law with basic components of cascades as stars and chains. As Influence propagates like SIS model in epidemiology, by understanding the deviations of patterns from these observed behaviors can help identify how the rumors or false information spreads online.

Each of the different types of approaches focused in this paper in terms of propagation of fake or untrustworthy information has multiple strengths and weaknesses which depends on the different scenarios and conditions. Multiple aspects like the user--related information, user’s content-related information, user’s social network-related information plays a key role in deciding how fake information or news spread rampantly on the online social networks.

**Recommendations & Future work**

**How to detect or identify fake accounts and predict friends vs. foes.**

The current fake account detection system is not sufficient to detect fake accounts and protect users from potentially harmful fake accounts. Future implementations of fake account detection systems should implement one of the other fake account detection systems.

I recommend the other fake account detection system such as “detection of clusters of fake accounts” as it is more time consuming and efficient with respect to the classification algorithms that they use. As day-in and day-out thousands of accounts will be created in OSN so it is recommended to detect the fake accounts at cluster level rather than account. The future work is to still reduce the complexity of machine learning pipeline such as combining two phases, cluster builder, profiler featurizer and make it single phase.

The systems developed so far do not yet completely eliminate the human intervention involved in the process of detecting fake or compromised accounts. Maybe future work can be perfected in a way that false positives and false negatives can be further reduced and a higher trust can be invested in automated systems to filter malicious accounts in OSNs.

**Studying the network and behavior of fake accounts**

More detection methods should aim at capturing fake accounts on a bigger scale and also in a timely manner, i.e. catch fake profiles when they get created, rather than at a later point of time when they have already been established.

OSNs grow at a very high rate in terms of profile counts, and considering this, a completely automated system to capture and remove fake accounts is best for the future. Manual interventions to analyse detection mechanism results are time consuming and prone to error.

Detection mechanisms should be able to cater to a variety of different social networks. A common representation model for online data has to be created where results coming from different networks can be analysed and actions taken.

In the results section we have our understanding on case study that privacy leakage of the legit users to the fake account users [8]. We discussed that the study shows the privacy leakage of only some users in online social networks. In near future, there is need to work on the large-scale set of users to find the self-disclosure to come up with more interesting results.

Coming to the understanding of Multiple Identity/Fake Account detection using non-verbal behavior [7], the method discussed uses Wikipedia as a tested online social network. Future work may need to study the method on different nonverbal behavior variables from different social media networks,. There may be also research to be done on combining both verbal and nonverbal behavior to the study to achieve the accuracy of predicting fake users.

We also have our understanding on Anomalous user behavior in Online Social Networks using PCA Analysis [9] mentioned in results section**.**  As discussed in the paper PCA analysis gives more systematic with general framework for studying the user behavior, there is still need of achieving good predictions and accuracy when combining with some supervised learning methods. Several research on how PCA analysis can be efficiently used on small scale crowdsourcing social networks should also to be considered in future work.

**What attributes lead to the success of fake accounts?**

The next generation tools discussed in [1] have methods to create fake accounts that are difficult to identify from genuine accounts based on the above discussed characteristics. The tools fetch information from different profiles and combine them to form a unique profile which are hard to be differentiated from real accounts.

There must be constant research on developing new tools and libraries that prevent generation of fake accounts. Google’s Recaptcha is one such library that resolves this issue for some time. Many parameters are taken into consideration before detecting a user as fake. The delay of clicking the button, cursor movements, browser, Ip address, browser history are taken into account before presenting a user with challenge. The challenge is not the simple captcha with a deformed text as there are tools that can resolve such captchas with high accuracy. Instead, different kinds of media are used like, identifying images, solving simple math questions, speech recognition, etc. This approach seems to perform well for time being until the bots come up with a new technique to bypass this. One such method is to mimic a real user like inducing delays for clicking etc. Also, people can be initially employed to solve this captchas and stored in database. The tool can use this knowledge and also adapt to learn using data mining techniques.

**How do people fall for fake accounts.**

In the first phase of this research study, we tried to understand the different factors that led to users falling for fake accounts. However, there are a multitude of reasons, and the attackers are quickly able to change perceived behavior to appear legitimate. Therefore, it becomes hard for both legitimate users and OSNs to discover which account is real or fake. In the second phase, we studied the different approaches OSNs could take in securing their networks, and what exactly they were looking for when identifying fake accounts. However, we realized that a lot of ways that OSNs discover fake accounts involves such a large data set and sometimes information that is not readily available to the average user. Thus, it was hard to always draw parallels between the research on OSN strategies and user safety.

For future work, we hope to bridge the gap by formulating best practices for the user based upon a specific type of network, the dating OSN. These networks have a lot of fake accounts and pose specific risks of identity theft, spamming, etc., similar to the risks mentioned on other social networks. By creating best practices for a certain OSN, and then seeing how it relates more generally, we can better understand the set of characteristics and behaviors that legitimate users should be guarded against. Even though this will not protect users completely from threats, it will encourage them be vigilant of the information they share, and how they communicate with strangers online.

**How does information diffuse on fake accounts compared to the real accounts?**

Despite the presence of numerous frameworks and methods to detect diffusion in social network, the performance of most of the trivial methods have been slow and by the time they are detected, the diffusion spreads deep into the network and hence the damage is done. In order to focus on early detection, we first try to understand how the information spreads on the network. Most of the current research is based on study focusing on a limited or closed group of individuals. In other words a subgraph is considered that may not replicate the original real graph. However in order to have more accuracy or impact, the subjects must include a diverse user base that will provide a broader understanding about how the information diffusion impacts them. During the detection and characterization of spammers, botnets the current system has very little information about the subject, hence leading to slower results, however if more views such as user history,age a much quicker detection of spams can be achieved and this improves the predictability of how much the information could diffuse into the system. Blacklisting of URLs encounters a problem due to the abundance of data delaying the detection, hence the focus should be on developing a quicker system for blocking these spammers or hindering the flow of information to stop the contagion in the initial stage. Also most of the research on this topic is domain specific, which makes it less impactful as compared to studying the global phenomenon of diffusion of information on social networks which is applicable for every domain. Further research on how and what information becomes viral, specific characteristics of URLs and pictures that tend to make them likeable, sentiment analysis is required to better predict how a particular information is going to impact the network. One important factor to focus is the effect of external factors while studying the contagion of untrustworthy information and how these factors can be handled in the initial phases itself to let the viral spread to breakdown.

**Evaluation Questions**

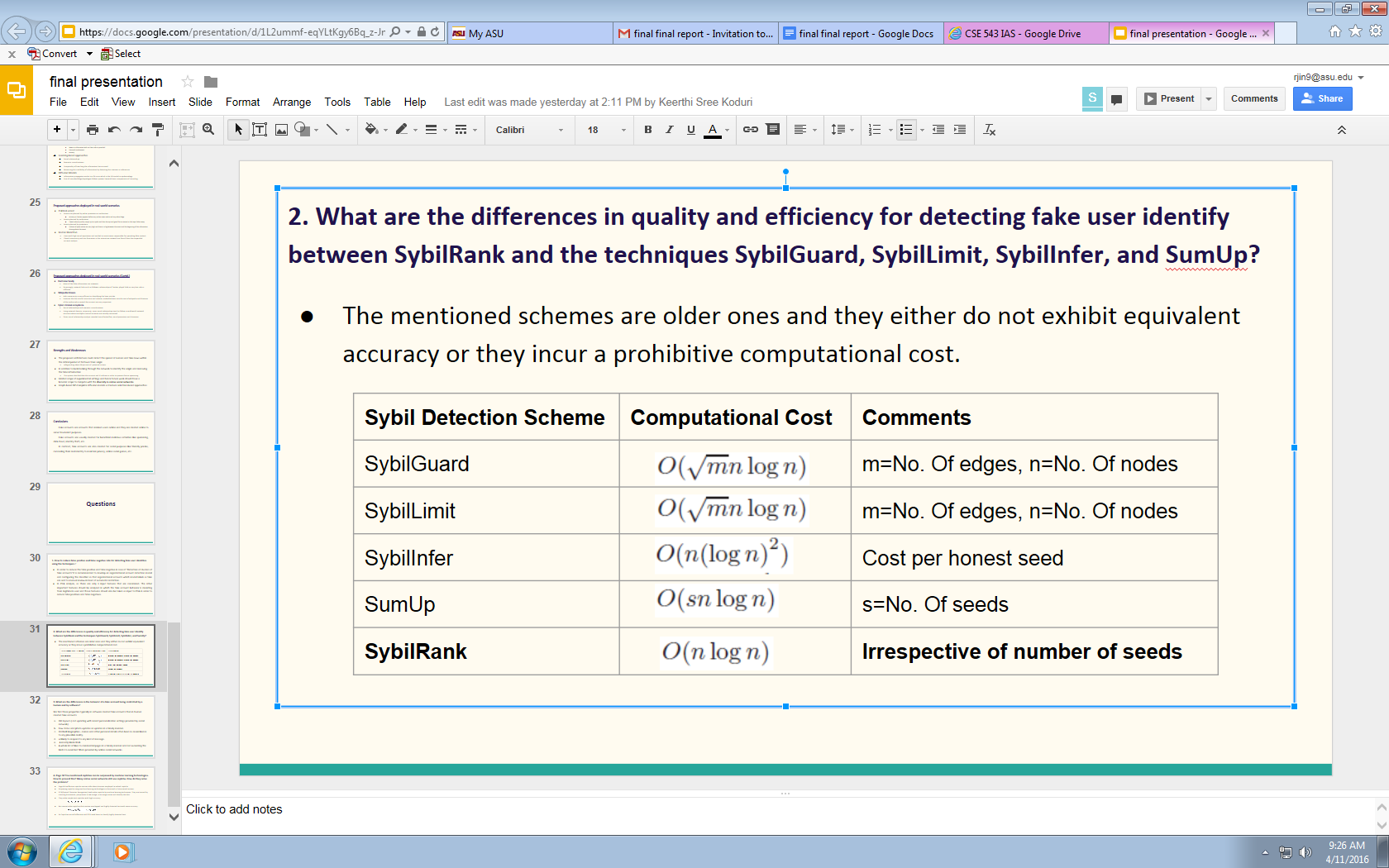
**1. How to reduce false positive and false negative rate for detecting fake user identities using the techniques ?**

In order to reduce the false positive and false negative in case of “Detection of clusters of fake accounts”it is recommended to develop an organizational account detection model and configuring the classifier so that organizational accounts which model labels as fake are sent to manual review instead of automatic restriction.

In PCA analysis, as there are only 3 input features that are considered. The other important features should be analyzed in which the fake account behavior is deviating from legitimate user and those features should also be taken as input to PCA in order to reduce false positives and false negatives.

**2. What are the differences in quality and efficiency for detecting fake user identify between SybilRank and the techniques SybilGuard, SybilLimit, SybilInfer, and SumUp?**

The mentioned schemes are older ones and they either do not exhibit equivalent accuracy or they incur a prohibitive computational cost.



**3. What are the differences in the behavior of a fake account being controlled by a human and by software?**

We find these properties typically in software created fake accounts than in human created fake accounts

* Old layouts ( not updating with recent personalization settings provided by social network).
* Few status and photo updates or updates in a timely manner.
* Oddball biographies - names and other personal details often bear no resemblance to any plausible reality.
* unlikely to respond to any kind of message.
* A mostly blank Wall.
* A whole lot of likes to commercial pages in a timely manner and not exceeding the limits to avoid bot filters provided by online social networks.

**4. Page 50 You mentioned captchas can be surpassed by machine learning technologies. How to prevent this? Many online social networks still use captcha. How do they solve the problem?**

* Page 50 Ineffective captcha section talks about Humans employed to solved captcha
* Surpassing captcha using machine learning technologies is discussed in Future work section.
* OCR(Optical Character Recognition) tools solve captcha by machine learning techniques. They are trained by scanning documents, pre-process a new image, scan image zones and identify the text.
* They solve simple text captchas with high accuracy

c1.png

* But cannot solve captchas that contain overlapped and highly distorted text with same accuracy.

c2.png

* So Captchas are still effective until OCR tools learn to classify highly distorted text.

**REFERENCES**

1. Hashmen Hefny, Ahmed El Azab, Amira M Idrees and Mahmoud A Mahmoud. **Fake Account Detection in Twitter Based on Minimum Weighted Feature set.** *World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering*.2015
2. Cao Xiao, David Mandell Freeman and Theodore Hwa**. Detecting Clusters of Fake Accounts in Online Social Networks**.In *Proceedings of 8th ACM workshop on Artifical Intelligence and security,*pages 91-101, New York, NY, USA, 2015. ACM.
3. Yazan Boshmaf, Ildar Muslukhov, Konstantin Beznosov, and Matei Ripeanu. 2011. **The socialbot network: when bots socialize for fame and money**. In *Proceedings of the 27th Annual Computer Security Applications Conference* (ACSAC '11). ACM, New York, NY, USA, 93-102.
4. Manuel Egele, Gianluca Stringhini, Christopher Kruegel , and Giovanni Vigna. **COMPA: Detecting Compromised Accounts on Social Networks.** University of California, Santa Barbara, Santa Barbara, CA
5. Qiang Cao, Michael Sirivianos, Xiaowei Yang, Tiago Pregueiro. **Aiding the Detection of Fake Accounts in Large Scale Social Online Services.** NSDI Conference 2012, San Jose.
6. A. Malhotra, L. Totti, W. Meira Jr., P. Kumaraguru, and V. Almeida. **Studying user footprints in diﬀerent online social networks**. In Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012).
7. M. Tsikerdekis and S. Zeadally. **Multiple account identity deception detection in social media using nonverbal behavior**. *IEEE TIFS*, 2014.
8. J. Ge, J. Peng and Z. Chen, "**Your Privacy Information are Leaking When You Surfing on the Social Networks: A Survey of the degree of online self- disclosure (DOSD)**," IEEE 13th Int’l Conf. on Cognitive Informatics & Cognitive Computing (ICCI\*CC’14), pp. 329-336, 2014.
9. B. Viswanath, M. A. Bashir, M. Crovella, S. Guha, K. P. Gummadi, B. Krishnamurthy, and A. Mislove. **Towards detecting anomalous user behavior in online social networks**. In *USENIX*, pages 223–238, Aug. 2014.
10. G. Kontaxis, I. Polakis, S. Ioannidis, and E. Markatos, “**Detecting social network proﬁle cloning**,” in PerCom, march 2011, pp. 295 –300.
11. L. Jin, H. Takabi, and J. B. Joshi. **Towards active detection of identity clone attacks on online social networks**. In Proceedings of the First ACM Conference on Data and Application Security and Privacy, CODASPY ’11.
12. Talel Abdessalem, Bogdan Cautis and Asma Souhli **Trust Management in Social Networks** Télécom ParisTech, LTCI – UMR CNRS 514146, rue Barrault – 75013 Paris – France, January 25, 2010
13. Yazan Boshmaf, Ildar Muslukhov, Konstantin Beznosov and Matei Ripeanu, **Key challenges in defending against malicious socialbots**, LEET'12 Proceedings of the 5th USENIX conference on Large-Scale Exploits and Emergent Threats, Pages 12-12.
14. Ehsan Ahmadizadeh,Erfan Aghasianb,Roohollah Fallah Nejad,Hossein Pour Taheric **An Automated Model to Detect Fake Profiles and botnets in Online Social Networks Using Steganography Technique** IOSR Journal of Computer Engineering e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 17, Issue 1, Ver. IV (Jan – Feb. 2015), PP 65-71
15. A. Pathak, “**An Analysis of Various Tools, Methods and Systems to Generate Fake Accounts for Social Media**,” Ph.D. dissertation, Northeastern University Boston, 2014.
16. Kumar, N. & Reddy, R. N. (2012) **Automatic Detection of Fake Profiles in Online Social Networks,** BTech Thesis.
17. Akshay J. Sarode and Arun Mishra. 2015. **Audit and Analysis of Impostors: An experimental approach to detect fake profile in online social network**. In *Proceedings of the Sixth International Conference on Computer and Communication Technology 2015* (ICCCT '15). ACM, New York, NY, USA, 1-8.
18. Kurt Thomas, Frank Li, Chris Grier, and Vern Paxson. 2014. **Consequences of Connectivity: Characterizing Account Hijacking on Twitter**. In Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security (CCS '14). ACM, New York, NY, USA, 489-50.
19. Arun Vishwanath. 2015. **Diffusion of deception in social media: Social contagion effects and its antecedents**. Information Systems Frontiers 17, 6 (December 2015), 1353-1367.
20. Chao Yang, Robert Harkreader, Jialong Zhang, Seungwon Shin, Guofei Gu.2012.**Analyzing Spammers Social Networks for Fun and Profit:a case study of cyber criminal ecosystem on twitter**. World Wide Web, pp.71-80(2012).
21. Gupta, Arpan, Hemank Lamba, and Ponnurangam Kumaraguru. "**$1.00 per rt #bostonmarathon #prayforboston: Analyzing fake content on twitter**." eCrime Researchers Summit (eCRS), 2013. IEEE, 2013.
22. Gupta, A., Lamba, H., Kumaraguru, P., Joshi, A.: **Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy**. In Proceedings of the 22nd International Conference on World Wide Web Companion, WWW 2013 Companion, 2013.
23. Srijan Kumar, Robert West and Jure Leskovec.2016.**Disinformation on the Web: Impact, Characteristics, and Detection of Wikipedia Hoaxes**. In 25th International World Wide Web Conference, 2016.
24. Jaeung Lee, Manish Agarwal, H.R.Rao .2015.**Message Diffusion through social network service: The case of rumor and non-rumor related tweets during Boston bombing 2013**. In Information Systems Frontiers, Volume 17 Issue 5 . pp 997-1005, 2015.
25. Jacob Ratkiewicz , Michael Conover , Mark Meiss , Bruno Gonçalves , Snehal Patil , Alessandro Flammini , Filippo Menczer, **Truthy: mapping the spread of astroturf in microblog streams**, Proceedings of the 20th international conference companion on World wide web, 2011.
26. W. Galuba, K. Aberer, D. Chakraborty, Z. Despotovic, and W. Kellerer. **Outtweeting the Twitterers - Predicting Information Cascades in Microblogs**. In 3rd Workshop on Online Social Networks, 2010
27. Chris Grier, Kurt Thomas, Vern Paxson, and Michael Zhang. 2010. **@spam: the underground on 140 characters or less**. In Proceedings of the 17th ACM conference on Computer and communications security (CCS '10). ACM, 27-37.
28. Yi-Min Wang, Ming Ma, Yuan Niu, and Hao Chen. 2007. **Spam double-funnel: connecting web spammers with advertisers**. In Proceedings of the 16th international conference on World Wide Web (WWW '07). ACM, New York, NY, USA, 291-300
29. Yinglian Xie, Fang Yu, Kannan Achan, Rina Panigrahy, Geoff Hulten, and Ivan Osipkov. 2008. **Spamming botnets: signatures and characteristics**. In Proceedings of the ACM SIGCOMM 2008 conference on Data communication (SIGCOMM '08), 171-182.
30. Marcelo Mendoza, Barbara Poblete, and Carlos Castillo. 2010. **Twitter under crisis: can we trust what we RT?**. In Proceedings of the First Workshop on Social Media Analytics (SOMA '10), 71-79
31. A. H. Wang, "**Don't follow me: Spam detection in Twitter**," Security and Cryptography (SECRYPT), Proceedings of the 2010 International Conference, 2010.
32. Anirudh Ramachandran and Nick Feamster. 2006. **Understanding the network-level behavior of spammers**. SIGCOMM Comput. Commun. Rev. 36, 4, 291-302.
33. Meredith Ringel Morris, Scott Counts, Asta Roseway, Aaron Hoff, and Julia Schwarz. “**Tweeting is believing?: understanding microblog credibility perceptions”** *In Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work (CSCW),* 2012
34. Adrien Guille, Hakim Hacid, Cecile Favre, and Djamel A. **Zighed. 2013. Information diffusion in online social networks: a survey.** *SIGMOD Rec.* 42, 2, July 2013.
35. Meeyoung Cha, Alan Mislove, and Krishna P. Gummadi. 2009. **A measurement-driven analysis of information propagation in the flickr social network.** In *Proceedings of the 18th international conference on World wide web* (WWW '09)
36. Lei Yang, Tao Sun, Ming Zhang, and Qiaozhu Mei. 2012. **We know what @you #tag**: does the dual role affect hashtag adoption?. In *Proceedings of the 21st international conference on World Wide Web* (WWW '12)
37. Nahon, K., Hemsley, J., Mason, R. M., Walker, S., & Eckert, J. (2013). **Information flows in events of political unrest.** iConference 2013 Proceedings (pp. 480-485)
38. Manuel Gomez Rodriguez, Jure Leskovec, and Andreas Krause. 2010. **Inferring networks of diffusion and influence**. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (KDD '10)
39. Karine Nahon, Jeff Hemsley, Shawn Walker and Muzammil Hussain, 2011, [**Fifteen Minutes of Fame: The Place of Blogs in the Life Cycle of Viral Political Informatio**](http://onlinelibrary.wiley.com/doi/10.2202/1944-2866.1108/abstract)**n**, Policy & Internet, Vol. 3(1)
40. Jure Leskovec, Mary McGlohon, Christos Faloutsos, Natalie Glance, Matthew Hurst.f **Cascading Behavior in Large Blog Graphs: Patterns and a Model** No. CMU-ML-06-113, 2006
41. Daniel M. Romero, Brendan Meeder, and Jon Kleinberg. 2011. twitter. **Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on** In *Proceedings of the 20th international conference on World wide web* (WWW '11)
42. Oh, O., Agrawal, M., & Rao, H. R. (2013). **Community intelligence and social media services: a rumor theoretic analysis of tweets during social crises**. MIS Quarterly, 37 (2), 407 – 426
43. Wael Jabr and Zhiqiang Zheng. 2014. **Know yourself and know your enemy: an analysis of firm recommendations and consumer reviews in a competitive environment**. *MIS Q.* 38, 3
44. Benjamin Doerr, Mahmoud Fouz, and Tobias Friedrich. 2012. **Why rumors spread so quickly in social networks**. *Commun. ACM* 55, 6 (June 2012), 70-75
45. Flavio Chierichetti, Silvio Lattanzi, and Alessandro Panconesi. 2011. **Rumor spreading in social networks.** *Theor. Comput. Sci.* 412, 24 (May 2011), 2602-2610
46. Noam Berger, Christian Borgs, Jennifer T. Chayes, and Amin Saberi. 2005. **On the spread of viruses on the internet.** In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms* (SODA '05). Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 301-310
47. Newman, M. E. J. (2002). **Spread of epidemic disease on networks**. *Physical Review E*, 66, 016128
48. T. Stein, E. Chen, and K Mangla. 2011. **Facebook Immune System**. Proceedings of the 4th Workshop on Social Network Systems (SNS ‘11). New York, NY, Article 8.
49. Vladislav Kontsevoi, Naim Lujan, and Adrian Orozco, **Detecting Subversion of Twitter** May 14, 2014.
50. Fabrıcio Benevenuto, Gabriel Magno, Tiago Rodrigues, and Virgılio Almeida, "**Detecting spammers on twitter,**" Collaboration, electronic messaging, anti-abuse and spam conference (CEAS). Vol. 6, 2010.
51. **The Fake project.** (Online). http://wafi.iit.cnr.it/theFakeProject/ (last retrieved on 30-10-2015).
52. Manish Gupta, Peixiang Zhao, and Jiawei Han, **Evaluating Event Credibility on Twitter Siam**, 2012.
53. L. Breiman, **Random forests Machine Learning**, 2001.
54. T. Joachims, **Learning to Classify Text Using Support Vector Machines:Methods, Theory, and Algorithms.** Boston: Kluwer Academic Publishers, 2002
55. Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze, **Introduction to Information Retrieval**. New York: Cambridge University, 2008
56. Manuel Fern\_andez Delgado, Eva Cernadas, Sen\_en Barro, and Dinani Amorim, **Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?**,Journal of Machine Learning Research, vol. 15, pp. 3133-3181, 2014.
57. Lior Rokach and Oded Maimon, **Data Mining and Knowledge Discovery Handbook** - Chapter 9 (Decision Trees), Oded Maimon and Lior Rokach, Eds., 2005.
58. David Kriesel, **A Brief Introduction to Neural Networks**.: dkriesel.com,2005. (Online) http://www.dkriesel.com/en/science/neural\_networks(last retrieved 30-10-2015).
59. B. E. Boser, I. M. Guyon, and V. N. Vapnik. **A training algorithm for optimal margin classifier**s. In Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory, pages 144{152. ACM Press, 1992.
60. N. Cristianini and J. Shawe-Taylor. **An Introduction to Support Vector Machines and Other Kernel-based Learning Methods.** Cambridge University Press, New York, NY, USA, 2000.
61. A. Rakotomamonjy. V**ariable selection using svm based criteria.** J. Mach. Learn. Res., 3:1357{1370, Mar. 2003.
62. V. N. Vapnik. **The Nature of Statistical Learning Theory.** Springer-Verlag New York, Inc., New York NY, USA, 1995.
63. Justin Chiu, Ajda Gokcen, Wenyi Wang, Xiaohua Yan. **Classification of Fake and Real Articles Based on Support Vector Machines** Language and Statistics -- Spring report 11761, 2013.
64. Tamotsu Shibutani. **Improvised News: A Sociological Study of Rumor**. An advanced study in sociology, 1966.